□ 现代统计学丛书

高等统计学概论

赵林城 王占锋 编著

GAODENG TONGJIXUE GAILUN

高等教育出版社·北京

图书在版编目（CIP）数据

高等统计学概论 / 赵林城，王占锋编著 . -- 北京 : 高等教育出版社，2016. 2
ISBN 978-7-04-044895-5

Ⅰ. ①高… Ⅱ. ①赵… ②王… Ⅲ. ①统计学 - 教材 Ⅳ. ① C8

中国版本图书馆 CIP 数据核字（2016）第 030449 号

策划编辑 王丽萍　责任编辑 李华英　封面设计 王凌波　版式设计 于 婕
责任校对 张小镝　责任印制 尤 静

出版发行 高等教育出版社
社　　址 北京市西城区德外大街4号
邮政编码 100120
印　　刷 北京佳信达欣艺术印刷有限公司
开　　本 787mm× 1092mm 1/16
印　　张 18
字　　数 370 千字
购书热线 010-58581118
咨询电话 400-810-0598
网　　址 http://www.hep.edu.cn
http://www.hep.com.cn
网上订购 http://www.landraco.com
http://www.landraco.com.cn
版　　次 2016 年 3月第 1 版
印　　次 2016 年 3月第1 次印刷
定　　价 69.00 元

本书如有缺页、倒页、脱页等质量问题，请到所购图书销售部门联系调换
版权所有 侵权必究
物 料 号 44895-00

版权.indd 1 2016-2-22 11:04:10

应用统计学丛书

书号	书名	著译者
	诊断医学中的统计学方法（第二版）	侯艳, 李康, 宇传华, 周晓华 译
	高等统计学概论	赵林城 王占锋 编著
	纵向数据分析方法与应用 （英文版）	刘宪
	生物数学模型的统计学基础 （第二版）	唐守正 李勇 符利勇 著
	R 软件教程与统计分析：入门到精通	潘东东 李启寨 唐年胜 译
	随机估计及 VDR 检验	杨振海
	随机域中的极值统计学：理论及应用(英文版)	Benjamin Yakir 著
	高等计量经济学基础	缪柏其、叶五一
	金融工程中的蒙特卡罗方法	Paul Glasserman 著 范韶华、孙武军 译
	大维统计分析	白志东、郑术蓉、姜丹丹
	结构方程模型：Mplus 与应用（英文版）	王济川、王小倩 著
	生存分析：模型与应用 (英文版)	刘宪
	MINITAB 软件入门：最易学实用的统计分析教程	吴令云 等 编著
	结构方程模型：方法与应用	王济川、王小倩、姜宝法 著
	结构方程模型：贝叶斯方法	李锡钦 著 蔡敬衡、潘俊豪、周影辉 译
	随机环境中的马尔可夫过程	胡迪鹤 著

续表

书号	书名	著译者
ISBN 978-7-04-025639-0	统计诊断	韦博成、林金官、解锋昌 编著
ISBN 978-7-04-025062-6	R 语言与统计分析	汤银才 主编
ISBN 978-7-04-024751-0	属性数据分析引论（第二版）	Alan Agresti 著 张淑梅、王睿、曾莉 译
ISBN 978-7-04-018293-4	金融市场中的统计模型和方法	黎子良、邢海鹏 著 姚佩佩 译

网上购书：www.hepmall.com.cn, www.gdjycbs.tmall.com, academic.hep.com.cn, www.china-pub.com, www.amazon.cn, www.dangdang.com

其他订购办法：

各使用单位可向高等教育出版社电子商务部汇款订购。书款通过支付宝或银行转账均可。购书免邮费，发票随书寄出（大批量订购图书，发票随后寄出）。

单位地址：北京西城区德外大街 4 号

电　　话：010-58581118　传　真：010-58581113

电子邮箱：gjdzfwb@pub.hep.cn

通过支付宝转账：

支付宝：gaojiaopress@sohu.com

名　称：高等教育出版社有限公司

通过银行转账：

户　　名：高等教育出版社有限公司

开 户 行：交通银行北京马甸支行

银行账号：110060437018010037603

前言

编写一本教科书, 断断续续, 前后经历了近八年的时间, 这在我还是第一次。退休前做科研, 完成一篇论文, 最短用几天时间, 最长的也不过一年。和陈希孺老师合作写过两本书, 也没费多长时间, 因为需要我写的是我们研究工作的一部分。教科书不一样, 总得写点与别人不一样的东西, 还得把难讲的东西写得简明易懂, 要做到这一点也颇费心思; 过去自己以为清楚了的东西, 等到下笔, 才发现仍有需要推敲的地方; 为了弄清一些问题的提法或者有关研究的来龙去脉, 还需翻阅若干参考文献。尤其不容易的是自己居然仍有这样的兴趣和耐心。

本书是为数理统计和相关专业的研究生或从事相关工作的人士撰写的。俗话说众口难调, 不敢奢望本书能适合所有人的需求, 但在作者心中, 还是希望本书能对朋友们在求知和求学方面有所裨益。本书着力于讨论数理统计的基本概念、理论、思想和方法, 在材料处理上有详有略, 有所侧重。随着统计科学和计量经济研究的发展, 为了处理一些复杂的统计模型, 越来越多的统计文献使用了过程收敛和经验过程的工具和方法。为了适应这种发展趋势, 使研究生尽早涉足统计研究的前沿, 本书在统计渐近理论方面加大篇幅讨论, 并着重介绍了经验过程理论在渐近统计中的应用。关于极值分布的渐近理论, 我们也尽可能引用了文献中最新的结果。

本书内容大致安排如下。作为统计理论体系的基础, 本书首先介绍了条件期望、条件概率和条件分布的理论, 以及统计决策理论的框架和有关概念。除了经典的点估计、假设检验、区间估计和 Bayes 统计外, 我们在第五章用较大篇幅介绍了统计渐近理论: 首先介绍了极限理论和渐近统计的基础知识, 以及 Cramér 关于似然方程解的经典理论; 由于 M-估计在统计应用中的重要地位, 以及经验过程理论在统计渐近理论中的主导作用和对未来的预期, 我们用了较大篇幅介绍如何借助于经验过程的理论和方法, 得出 M-估计的渐近分布, 进而建立 Le Cam 关于极大似然估计渐近分布的一种精妙方法; 而有关的过程收敛和经验过程的预备知识可以在附录中找到。之所以要作这样的努力, 不在于介绍多少具体的知识, 而是希望通过展现不同的视

角和处理方法, 使读者能在无形中提高一种内在的能力。

本书选取了一定数量的例题, 有些是作为理论或定理的应用, 有些则是理论的延伸。一部分例题采用了独特的解法, 希望能起一些示范作用。为了提高学生的解题能力, 每章最后都附有适量的习题, 除了供读者自我训练之外, 有些则是正文的补充。较难的习题前加了 * 号, 一些习题则加上了提示。为了提高自身解决问题的能力, 希望有志于这一学科的读者除了做好基本的习题之外, 也要能解一部分难题, 通过努力, 使自己的解题能力再上一个台阶。作者切身的体会是, 在上学时自己解出一些难题, 不仅会提高自己的信心和兴趣, 而且也是一种让人久久难忘的经历。无论将来从事科学研究、教学还是实际应用工作, 不断打好基础也许会成为一生的功课。

在本书写作过程中, 部分内容已经给研究生讲了两次, 每次 80 学时, 实践看来这些学时仍然不够, 讲授时可以在内容上有所取舍。初学时有些较难的内容可以略去, 较难的证明也可以不讲。关于 M-估计渐近理论的讲述, 可以限制在经验过程有关方法和结果的应用层面, 而不涉及经验过程的理论研究本身。有关的专门理论可以鼓励学生自己钻研, 或者另外组织讨论班仔细研读。作者过去就曾这样做过。

本书在正文或习题中多处引用了陈希孺 (1981) (1999), Lehmann (1983) (1986), 茆诗松, 王静龙, 濮晓龙 (1998), van der Vaart (1998), Reiss (1989), Ferguson (1996) 以及 Shao (2005) 等著作中的有关材料, 有些已在引用处注明, 但遗漏之处在所难免, 除在此作统一说明之外, 还望读者和有关方面谅解。

第四章区间估计是本书另一作者王占锋撰写的。其中信仰推断部分引进了杨振海等人采用的垂直密度表示的内容, 并在内容和篇幅上作了微型化处理, 以便让读者用尽量少的时间了解这一方法的概貌。本书其余各章和附录部分为本人撰写。为使风格一致, 统筹工作也由本人完成。

本书在写作过程中, 得到了缪柏其、韦来生、吴耀华、杨亚宁、陈敏诸位教授以及张伟平、胡治水副教授等人的鼓励和实际帮助。陈子沐同学帮助作者解决了绘图以及录入中遇到的一些疑难问题。在此向他们一并致谢。

由于作者水平所限, 本书错误和疏漏之处在所难免, 还望读者不吝指教。

赵林城

2015 年 5 月

目录

第一章　基本概念 **1**
1.1　条件期望和条件概率 1
1.1.1　与可测变换有关的两个定理 1
1.1.2　其他有关的预备定理 3
1.1.3　条件期望的定义和性质 5
1.1.4　条件概率的定义和性质 7
1.1.5　条件概率分布 7
1.2　样本空间与分布族 10
1.2.1　样本空间与样本分布族 11
1.2.2　指数型分布族 12
1.2.3　若干常用分布族 16
1.3　统计推断与统计决策理论的基本概念 18
1.3.1　统计推断 18
1.3.2　统计决策问题的三个要素 19
1.3.3　统计决策函数及其风险函数 20
1.4　统计量 23
1.4.1　定义和例子 23
1.4.2　与正态样本有关的抽样分布 25
1.4.3　对称幂等方阵与 χ^2 分布 28
1.5　充分统计量 28
1.5.1　定义和例子 29
1.5.2　因子分解定理 30
1.5.3　充分性原则 33

1.6 分布族的完全性和完全统计量 · · · · · 35
1.6.1 基本概念, Basu定理 · · · · · 35
1.6.2 一些常见分布族的完全统计量 · · · · · 36
1.7 凸损失函数 · · · · · 40
1.8 习题和补充 · · · · · 42

第二章 点估计 · · · · · 48
2.1 无偏估计 · · · · · 49
2.1.1 风险一致最小的无偏估计 · · · · · 49
2.1.2 Cramér-Rao 不等式 · · · · · 56
2.1.3 多个参数的情况 · · · · · 59
2.2 同变估计 · · · · · 61
2.2.1 同变性概论 · · · · · 61
2.2.2 风险一致最小的同变估计 · · · · · 63
2.3 Bayes 估计 · · · · · 68
2.3.1 Bayes 统计决策的基本框架 · · · · · 68
2.3.2 一些重要情形的 Bayes 估计 · · · · · 71
2.3.3 共轭先验分布族 · · · · · 74
2.3.4 广义 Bayes 估计 · · · · · 76
2.3.5 经验 Bayes 估计 · · · · · 79
2.3.6 关于 Bayes 统计推断的一些说明 · · · · · 83
2.3.7 先验分布的选取, 无信息先验分布 · · · · · 87
2.4 Minimax 估计 · · · · · 90
2.5 估计的容许性 · · · · · 94
2.6 习题和补充 · · · · · 100

第三章 假设检验 · · · · · 106
3.1 基本概念 · · · · · 106
3.1.1 统计假设和检验函数 · · · · · 106
3.1.2 假设检验问题的 Neyman-Pearson 提法 · · · · · 107
3.2 一致最优检验 · · · · · 108
3.2.1 Neyman-Pearson 基本引理 · · · · · 108
3.2.2 单调似然比分布族与 UMP 检验 · · · · · 110
3.2.3 假设检验与两决策问题 · · · · · 113

3.3 Neyman-Pearson 基本引理的推广 · · · · · 114
3.4 无偏检验 · · · · · 116
3.4.1 检验的无偏性 · · · · · 116
3.4.2 单参数指数族的 UMP 无偏检验 · · · · · 117
3.4.3 多参数指数族的 UMP 无偏检验 · · · · · 119
3.4.4 与正态有关的检验 · · · · · 122
3.5 不变检验 · · · · · 125
3.5.1 问题的提法 · · · · · 125
3.5.2 一致最优不变检验 · · · · · 126
3.6 习题和补充 · · · · · 128

第四章 区间估计 · · · · · 135
4.1 基本概念 · · · · · 135
4.2 构建区间估计的方法 · · · · · 137
4.2.1 枢轴变量法 · · · · · 137
4.2.2 基于连续随机变量构建置信区间 · · · · · 139
4.2.3 基于离散随机变量构建置信区间 · · · · · 140
4.2.4 假设检验法 · · · · · 143
4.2.5 大样本方法 · · · · · 144
4.3 区间估计的优良性 · · · · · 145
4.4 Bayes 区间估计 · · · · · 149
4.5 信仰推断法 · · · · · 152
4.6 习题和补充 · · · · · 157

第五章 统计渐近理论 · · · · · 162
5.1 估计的相合性和渐近正态性 · · · · · 163
5.1.1 基本概念 · · · · · 163
5.1.2 Delta 方法 · · · · · 165
5.1.3 矩估计 · · · · · 166
5.2 极大似然估计 · · · · · 169
5.2.1 一般概念 · · · · · 169
5.2.2 指数族情形的 MLE · · · · · 172
5.2.3 MLE 的渐近理论 · · · · · 174
5.3 M-估计 · · · · · 179
5.3.1 M-估计的概念 · · · · · 179
5.3.2 M-估计的相合性 · · · · · 181

5.3.3 M-估计的收敛速度 · · · · · 182
5.3.4 M-估计的渐近正态性 · · · · · 184
5.3.5 再访 MLE · · · · · 188
5.4 契合性 · · · · · 194
5.5 大样本检验 · · · · · 199
5.5.1 似然比检验 · · · · · 199
5.5.2 拟合优度检验 · · · · · 203
5.6 次序统计量 · · · · · 213
5.6.1 基本概念 · · · · · 214
5.6.2 次序统计量的极限分布 · · · · · 218
5.6.3 极值分布的参数估计 · · · · · 224
5.6.4 L-统计量 · · · · · 228
5.7 从最小二乘谈起 · · · · · 230
5.7.1 线性回归的最小二乘估计 · · · · · 230
5.7.2 线性模型中的 M-估计 · · · · · 231
5.7.3 广义线性模型 · · · · · 234
5.7.4 其他回归模型 · · · · · 235
5.8 习题和补充 · · · · · 238

附录 · · · · · **245**
A.1 检验函数空间的一个弱紧性定理 · · · · · 245
A.2 随机变量序列的各种收敛性 · · · · · 247
A.3 距离空间上的随机元序列的收敛性 · · · · · 250
A.4 经验过程 · · · · · 254
A.4.1 经验分布 · · · · · 254
A.4.2 极大不等式 · · · · · 257
A.4.3 随机函数 · · · · · 257
A.4.4 $\mathcal{F}$ 改变为 $\mathcal{F}_n$ 的情形 · · · · · 258

参考文献 · · · · · **261**

索引 · · · · · **271**

第一章

基本概念

1.1 条件期望和条件概率

在数理统计的理论体系中, 条件期望和条件概率的理论占有重要的地位. 本节将以适合于统计需要的形式, 介绍这一理论的一些基本事实. 这些内容都是读者应当掌握的.

为了理论发展的需要, 下面先介绍一些重要的预备定理.

1.1.1 与可测变换有关的两个定理

在以下的陈述中, 空间 $\mathcal{X}$ 是具有某种性质的元素 (称为点) 构成的一个非空集合, $\mathcal{B}$ 为 $\mathcal{X}$ 的某些子集构成的一个 σ-域, 此时称 $(\mathcal{X},\mathcal{B})$ 为一**可测空间**, $\mathcal{B}$ 中的集合称为可测集. 最为常见的可测空间是 $(\mathbf{R}^k,\mathcal{B}_k)$, 其中 $\mathbf{R}^k$ 是一 k 维欧氏空间, $\mathcal{B}_k$ 是由 $\mathbf{R}^k$ 中的开集系生成的 σ-域, 即包含所有开集的最小 σ-域, 通常把这个 σ-域称为 Borel-域, 其中的集合则称为 Borel 集.

考虑两个非空集合 $\mathcal{X}$ 和 $\mathcal{T}$. 集合 $\mathcal{X}$ 到集合 $\mathcal{T}$ 内的一个变换 (或函数) T 可记为 $T:\mathcal{X}\mapsto\mathcal{T}$. 定义于 $\mathcal{X}$ 上取值于 $\mathbf{R}'=(-\infty,\infty)$[①] 的函数称为 $\mathcal{X}$ 上的实值函数. 对于 $\mathcal{T}$ 的任一子集 G 和它的任一子集类 $\mathcal{G}$, 分别称 $T^{-1}(G):=\{x\in\mathcal{X}:T(x)\in G\}:=\{T\in G\}$ 和 $T^{-1}(\mathcal{G}):=\{T^{-1}(G):G\in\mathcal{G}\}$ 为集合 G 和集类 $\mathcal{G}$ 的原像.

由于 T^{-1} 保持集合的运算和包含关系, 所以一个域的原像为一个域, σ-域的原像为一个 σ-域.

设 $T:\mathcal{X}\mapsto\mathcal{T}$ 为可测空间 $(\mathcal{X},\mathcal{B})$ 到可测空间 $(\mathcal{T},\mathcal{G})$ 内的一个变换, 如果对任

[①] 实数集 $\mathbf{R}'$ 也可简记为 $\mathbf{R}$, 而广义直线 $\overline{\mathbf{R}}'$ 即可简记为 $\overline{\mathbf{R}}$

— $G \in \mathcal{G}$, $T^{-1}(G) \in \mathcal{B}$, 则称 T 是 $\mathcal{B}|\mathcal{G}$ 可测的, 这一条件等价于 $T^{-1}(\mathcal{G}) \subset \mathcal{B}$. 此时 $T^{-1}(\mathcal{G})$ 为 $\mathcal{B}$ 的一个子 σ-域. $(\mathcal{X}, \mathcal{B})$ 到空间 $(\mathbf{R}^k, \mathcal{B}_k)$ 内的可测函数, 常简称为 $\mathcal{B}$-可测函数. 在没有歧义时涉及的 σ-域则无需交代.

设 μ 为 $\mathcal{B}$ 上的一个测度, T 为 $(\mathcal{X}, \mathcal{B})$ 到 $(\mathcal{T}, \mathcal{G})$ 内的一个可测变换. 对于每个 $G \in \mathcal{G}$, 定义 $\mu^*(G) = \mu(T^{-1}(G))$. 不难证明 μ^* 为 $\mathcal{G}$ 上的一个测度, 称为 μ 在 T 之下的导出测度. 注意 μ 和 μ^* 是定义在不同的 σ-域上的测度. 在测度论和概率论中, 称 $(\mathcal{X}, \mathcal{B}, \mu)$ 为一**测度空间**, 而当 μ 为一概率测度时, 则称其为**概率空间**.

设 B 为 $\mathcal{X}$ 的任一子集, 其余集 $\mathcal{X} \setminus B$ 记为 B^c. 集合 B 的示性函数 I_B 定义为

$$I_B(x) = \begin{cases} 1, & \text{当 } x \in B, \\ 0, & \text{当 } x \in B^c. \end{cases}$$

集合 B 的示性函数有时也记为 $I(B)$.

随机元和样本空间是数理统计中的基本概念. 统计问题研究的是带有随机性的数据, 这些数据可以用一个随机元 X 来表示. 在数理统计中, 空间 $\mathcal{X}$ 上的一个**随机元** X 即为有关的概率空间 $(\Omega, \mathcal{A}, P)$ 到可测空间 $(\mathcal{X}, \mathcal{B})$ 内的一个可测映照. 作为随机元 X 取值的空间, $(\mathcal{X}, \mathcal{B})$ 称为**样本空间**. 一个重要的情形是上述定义中的空间 $\mathcal{X}$ 为一距离空间, $\mathcal{B}$ 为它的 Borel σ-域, 即由 $\mathcal{X}$ 中的开集系生成的 σ-域. 在许多常见的统计问题中, X 为一有限维随机向量, 在本书中, $\mathcal{X}$ 常常取为 $\mathbf{R}^k$ 或它的一个适当的 Borel 子集, 连同相应的 Borel-域 $\mathcal{B}_k$, 称这样的样本空间是**欧氏的**.

以后我们将会看到, 一个统计问题的表述, 一般不再提及有关的概率空间 $(\Omega, \mathcal{A}, P)$, 样本空间及由随机元导出的概率分布族即为统计研究的出发点.

需要指出的是, 仅有欧氏样本空间是远远不够的, 函数型数据和过程统计的研究, 就要涉及可距离化样本空间.

下面引出两个重要引理.

引理 1.1 *设 T 为可测空间 $(\mathcal{X}, \mathcal{B})$ 到可测空间 $(\mathcal{T}, \mathcal{G})$ 的可测变换, $\mathcal{B}_0 := T^{-1}(\mathcal{G})$ 为由 T 导出的 $\mathcal{B}$ 的子 σ-域, 则一个实值 $\mathcal{B}$-可测函数 f 为 $\mathcal{B}_0$-可测的充要条件为: 存在一个 $\mathcal{G}$-可测函数 g, 使 $f(x) = g(T(x))$ 对所有 $x \in \mathcal{X}$ 成立.*

证明 充分性显然. 事实上, 假定这样的函数 g 存在, 则对任一 Borel 集 C,

$$f^{-1}(C) = T^{-1}\big(g^{-1}(C)\big) \in T^{-1}(\mathcal{G}) = \mathcal{B}_0,$$

故 f 为 $\mathcal{B}_0$-可测.

为证必要性, 设 f 为 $\mathcal{B}_0$-可测, 则对固定的 n, 集合

$$B_{in} = \{x : i/2^n < f(x) \leqslant (i+1)/2^n\}, \quad i = 0, \pm 1, \pm 2, \cdots$$

为互斥的 $\mathcal{B}_0$ 集, 其并集为 $\mathcal{X}$, 因而存在 $G_{in} \in \mathcal{G}$, 使 $B_{in} = T^{-1}(G_{in})$. 不难证明, 存在 G_{in} 的子集 $G_{in}^* \in \mathcal{G}$, 使得对固定的 n, $G_{in}^*, i = 0, \pm 1, \pm 2, \cdots$ 两两不交, 且满足 $B_{in} = T^{-1}(G_{in}^*)$. 实际上, 只需取 $G_{in}^* = G_{in} \setminus \bigcup_{j \neq i} G_{jn}$ (请读者补出证明的细节).

定义

$$f_n(x) = i/2^n, \quad \text{如果 } x \in B_{in}, \quad i = 0, \pm 1, \pm 2, \cdots,$$

$$g_n(t)=\begin{cases} i/2^n, & \text{如果 } t\in G_{in}^*,\ i=0,\pm1,\pm2,\cdots,\\ 0, & \text{否则,}\end{cases}$$

那么 $f_n(x)=g_n(T(x)), x\in\mathcal{X}$. 因为函数 g_n 为 $\mathcal{G}$-可测, 故使得 $g_n(t)$ 收敛于有限极限的点 t 所成的集合 $V\in\mathcal{G}$. 令 $S=T(\mathcal{X})$ 为 T 的值域, 则对 $t\in S$, 存在 $x\in\mathcal{X}$, 使 $t=T(x)$, 故

$$\lim g_n(t)=\lim g_n[T(x)]=\lim f_n(x)=f(x)$$

对所有 $x\in\mathcal{X}$ 成立, 因而 $S\subset V$. 定义

$$g(t)=\begin{cases}\lim g_n(t), & \text{当 } t\in V,\\ 0, & \text{否则,}\end{cases}$$

则 g 就具有所要的性质. ■

参看 Doob (1953) 和 Lehmann (1986). 严加安 (1981) 和严士健, 刘秀芳 (1994) 提供了另一种证明方法. 上述定理称为 **Doob 复合函数可测性定理**, 它也可等价地表述如下.

设 T 为 $\mathcal{X}$ 到 $(\mathcal{T},\mathcal{G})$ 内的一个变换, $\mathcal{B}:=T^{-1}(\mathcal{G})$ 为 $\mathcal{X}$ 上由 T 导出的 σ-域, 则 $\mathcal{X}$ 上的一个实值函数 f 为 $\mathcal{B}$-可测的充要条件为: 存在一个 $\mathcal{G}$-可测函数 g, 使 $f(x)=g(T(x))$ 对所有 $x\in\mathcal{X}$ 成立.

上述函数 f 和 g 的积分之间的关系由下面的引理给出.

引理 1.2 (积分变换定理) 设 T 为 $(\mathcal{X},\mathcal{B})$ 到 $(\mathcal{T},\mathcal{G})$ 的可测变换, μ 为 $\mathcal{B}$ 上的一个测度, μ^* 为其导出测度, 则对任何 $\mathcal{G}$-可测的实值函数 $g(t)$ 和任一 $G\in\mathcal{G}$, 有

$$\int_{T^{-1}(G)} g[T(x)]d\mu(x)=\int_G g(t)d\mu^*(t). \tag{1.1}$$

上式成立是指只要一边的积分存在, 另一边的积分也存在, 并且两者相等.

证明 不失一般性, 不妨假定 G 就是整个空间 $\mathcal{T}$. 若 $g=I_{G^*}$ 为集合 G^* 的示性函数, $G^*\in\mathcal{G}$, 则 (1.1) 的左右两边分别简化为 $\mu\big(T^{-1}(G^*)\big)$ 和 $\mu^*(G^*)$, 由 μ^* 的定义, 引理显然成立. 因而 (1.1) 对简单函数, 对非负可测函数, 继而对积分存在的所有函数都成立. ■

1.1.2 其他有关的预备定理

首先介绍著名的 Radon-Nikodym 定理. 为此, 设 μ 和 ν 为可测空间 $(\mathcal{X},\mathcal{B})$ 上的两个测度. 若对任何 $B\in\mathcal{B}$,

$$\mu(B)=0\Rightarrow\nu(B)=0,$$

则称 ν 关于 μ 绝对连续, 或称测度 ν 受控于测度 μ, 记为 $\nu\ll\mu$. 如果 $\nu\ll\mu$, 且 $\mu\ll\nu$, 则称这两个测度等价. 这些定义容易推广到 μ 为 $(\mathcal{X},\mathcal{B})$ 上的有号测度的情形.

称 μ 为可测空间 $(\mathcal{X},\mathcal{B})$ 上的一个 σ-**有限测度**, 如果存在一列 μ-测度有限的可

测集 A_n, 使 $\mathcal{X}=\bigcup_{n=1}^{\infty}A_n$. 定义

$$\nu(B)=\sum_{n=1}^{\infty}\frac{\mu(A_nB)}{1+2^n\mu(A_n)},\quad B\in\mathcal{B},$$

则 ν 与 μ 等价, 且 ν 为有限测度.

定理 1.3 (Radon-Nikodym 定理) 设 μ 和 ν 为可测空间 $(\mathcal{X},\mathcal{B})$ 上的两个 σ-有限测度, 则 ν 关于 μ 绝对连续的充要条件为: 存在一个几乎处处有限的非负的 $\mathcal{B}$-可测函数 f, 使

$$\nu(B)=\int_B f(x)d\mu(x),\quad \text{对任何 } B\in\mathcal{B}. \tag{1.2}$$

如果不计其在一个 μ-零测集上的取值, 则 f 由关系式 (1.2) 唯一确定.

更一般地, 上述定理也适用于 ν 为 $(\mathcal{X},\mathcal{B})$ 上的 σ-有限的有号测度的情形.

此时称 f 为 ν 关于 μ 的 Radon-Nikodym 导数, 记为

$$d\nu=fd\mu,\quad \text{或}\quad f=d\nu/d\mu.$$

在 ν 为概率测度的情形, f 称为 ν 关于 μ 的概率密度. 有时为了指出上述 f 确定到不计它在一个 μ-零测集上的取值, 常记为

$$f=d\nu/d\mu,\quad \text{a.e. } \mu\quad \text{或}\quad f=d\nu/d\mu,\quad \text{a.e. } (\mathcal{B},\mu).$$

此处我们约定, 如果一个命题 $Q(x)$ 对除去一个 μ-零测集的所有 x 成立, 则称 $Q(x)$ a.e. μ 成立; 如要进一步强调 μ 的定义域 $\mathcal{B}$, 则写成 $Q(x)$, a.e. $(\mathcal{B},\mu)$. 设 $\mathcal{P}$ 为一概率测度族, 若对于每个 $P\in\mathcal{P}$ 都有 $Q(x)$, a.e. P, 则称 $Q(x)$ a.e. $\mathcal{P}$ 成立.

在此我们顺便指出, 在证明上述定理中 f 的唯一性时用到一个有用的事实: 设 f 和 g 是两个 $\mathcal{B}$-可测且 μ-可积的函数, 使 $\int_B f(x)d\mu(x)=\int_B g(x)d\mu(x)$ 对任何 $B\in\mathcal{B}$ 成立, 则 $f=g$, a.e. μ. 事实上, 令 $B=\{f>g\}$, 则 $B\in\mathcal{B}$, 由假定,

$$\int_{\{f>g\}}[f(x)-g(x)]d\mu(x)=0,$$

因而 $\mu(f>g)=0$. 同样也有 $\mu(f<g)=0$.

需要介绍的下一个定理是关于重积分的 Fubini 定理. 设 B 和 C 为两个集合, 集合 $B\times C:=\{(x,y):x\in B,y\in C\}$ 称为 B 和 C 的直积 (或 Cartesian 积). 设 $(\mathcal{X},\mathcal{B})$ 和 $(\mathcal{Y},\mathcal{C})$ 为两个可测空间, $\mathcal{B}\times\mathcal{C}$ 是包含形如 $B\times C(B\in\mathcal{B},C\in\mathcal{C})$ 的所有集合的最小 σ-域. 设 μ 和 ν 分别为 $(\mathcal{X},\mathcal{B})$ 和 $(\mathcal{Y},\mathcal{C})$ 上的 σ-有限测度, 则存在 $(\mathcal{X}\times\mathcal{Y},\mathcal{B}\times\mathcal{C})$ 上唯一的测度 $\lambda=\mu\times\nu$, 称为 μ 和 ν 的乘积测度, 使得对所有的 $B\in\mathcal{B},C\in\mathcal{C}$,

$$\lambda(B\times C)=\mu(B)\nu(C).$$

在可测性方面, 以下事实值得一提: 设 $f(\cdot,\cdot)$ 为 $\mathcal{B}\times\mathcal{C}$-可测函数, 则对 $\forall x\in\mathcal{X}$, $f(x,\cdot)$ 为一 $\mathcal{C}$-可测函数, 而对 $\forall y\in\mathcal{Y}$, $f(\cdot,y)$ 为一 $\mathcal{B}$-可测函数.

定理 1.4 (Fubini 定理) 设 μ 和 ν 分别为 $(\mathcal{X},\mathcal{B})$ 和 $(\mathcal{Y},\mathcal{C})$ 上的 σ-有限测度, $\lambda=\mu\times\nu$. 设 $f(x,y)$ 关于 λ 可积, 则

(i) 按测度 ν 对几乎所有的 y, $f(\cdot,y)$ 关于 μ 可积;

(ii) 函数 $\int_{\mathcal{X}} f(x,y)d\mu(x)$ 关于 ν 可积, 且

$$\int_{\mathcal{X}\times\mathcal{Y}} f(x,y)d\lambda(x,y)=\int_{\mathcal{Y}}\left[\int_{\mathcal{X}} f(x,y)d\mu(x)\right]d\nu(y).$$

如果积分区域是全空间, 今后常常将其略去不写.

1.1.3 条件期望的定义和性质

设 $(\mathcal{X},\mathcal{B},P)$ 为一概率空间, T 为 $(\mathcal{X},\mathcal{B})$ 到 $(\mathcal{T},\mathcal{G})$ 的可测变换, $\mathcal{B}_0:=T^{-1}(\mathcal{G})$ 为由 T 导出的 $\mathcal{B}$ 的子 σ-域. 在统计中, P 为随机元 X (也可视为一个样本) 服从的概率分布, T 为一个统计量. 设 f 是一个 $(\mathcal{B},P)$-可积 (即 $\mathcal{B}$-可测且 P-可积) 的实值函数, 则 $\nu(B):=\int_B f dP$ 对所有 $B\in\mathcal{B}$ 有定义, 因而也对所有 $B\in\mathcal{B}_0$ 有定义. 局限于 σ-域 $\mathcal{B}_0$ 上, $\nu\ll P$, 由 Radon-Nikodym 定理, 存在 $(\mathcal{B}_0,P)$-可积函数 f_0, 使

$$\int_{B_0} f dP=\int_{B_0} f_0 dP,\quad \text{对所有}\ \ B_0\in\mathcal{B}_0, \tag{1.3}$$

此处, f_0 由上式 a. e. $(\mathcal{B}_0,P)$ 唯一确定. 定义 T 的概率分布 P^* 如下: $P^*(G)=P(T^{-1}(G)),G\in\mathcal{G}$. 由引理 1.1, 存在一个 $\mathcal{G}$-可测函数 g, 使 $f_0(x)=g(T(x))$ 对所有 $x\in\mathcal{X}$ 成立. 由引理 1.2, 函数 g 可直接借助于 f 由

$$\int_{T^{-1}(G)} f(x)dP(x)=\int_G g(t)dP^*(t),\quad \text{对所有}\ G\in\mathcal{G} \tag{1.4}$$

定义. 如果不计其在一个 P^*-零测集上的取值, 则 g 是唯一确定的.

定义 1.1 设 f_0 为满足 (1.3) 的任一 $\mathcal{B}_0$-可测函数, 则称 $f_0(X)$ 为在给定子 σ-域 $\mathcal{B}_0$ 或统计量 $T(X)$ 的条件下 $f(X)$ 的**条件期望**, 记为

$$f_0(X)=\mathrm{E}[f(X)|\mathcal{B}_0]=\mathrm{E}[f(X)|T(X)],$$

或

$$f_0(x)=\mathrm{E}[f(X)|T(x)]:=\mathrm{E}[f(X)|T(X)=T(x)].$$

或者等价地, 称 $g(T)$ 为在给定 T 的条件下 $f(X)$ 的**条件期望**, 其中 g 为满足 (1.4) 的任一 $\mathcal{G}$-可测函数, 记为

$$g(T)=\mathrm{E}[f(X)|T],\ \text{或者}\ g(t)=\mathrm{E}[f(X)|t]:=\mathrm{E}[f(X)|T=t].$$

需要特别指出的是, 条件期望的上述定义只是存在性的, 而不是构造性的, 没有提供任何计算方法. 为求解条件期望, 一般是先借助直观 "找" 出一个解, 再用上述定义加以验证. 对初学者而言, 这就是困难之所在, 只有多做练习才能逐步把握它.

例 1.1 设 $X_1,\cdots,X_n$ 是独立同分布 (i.i.d.) 的随机变量, 具有相同概率分布 P 且其分布函数连续. 令

$$T(x_1,\cdots,x_n)=(x_{(1)},\cdots,x_{(n)}),$$

其中 $x_{(1)}\leqslant x_{(2)}\leqslant\cdots\leqslant x_{(n)}$ 指出这些 x_i 之间的次序. 称 $(X_{(1)},\cdots,X_{(n)})$ 为 $X_1,\cdots,X_n$ 的次序统计量, 或者称其为从分布 P 中抽取的容量为 n 的次序样本.

不失一般性, 我们可以只限于考虑 $x_{(1)} < x_{(2)} < \cdots < x_{(n)}$ 的那些点, 因为有两个坐标相等的概率为 0. 因而 $\mathcal{X}$ 可取为不同坐标的所有 n 数组的集合, $\mathcal{T}$ 为所有有序 n 数组的集合, $\mathcal{B}$ 和 $\mathcal{G}$ 分别为 $\mathcal{X}$ 和 $\mathcal{T}$ 中 Borel 子集的类. 在 T^{-1} 之下, 由单个向量 $a = (a_1, \cdots, a_n)$ 组成的单点集变换成由 $n!$ 个点组成的集合, 它可通过将 a 的坐标进行所有可能的置换而得到. 因此, $\mathcal{B}_0$ 是所有对称的 Borel 子集 B_0 组成的集类, 对称的含义为, 如果 B_0 包含点 $x = (x_1, \cdots, x_n)$, 则它也包含所有形如 $(x_{i_1}, \cdots, x_{i_n})$ 的点, 此处 $(i_1, \cdots, i_n)$ 是 $(1, \cdots, n)$ 的任一置换. 对于任何可积函数 f, 令

$$f_0(x) = \frac{1}{n!} \sum f(x_{i_1}, \cdots, x_{i_n}),$$

此处, $(i_1, \cdots, i_n)$ 跑遍 $(1, \cdots, n)$ 的所有置换. 函数 f_0 是 $\mathcal{B}_0$-可测的, 因为它关于 n 个自变量是对称的. 由于

$$\int_{B_0} f(x_1, \cdots, x_n) dP(x_1) \cdots dP(x_n) = \int_{B_0} f(x_{i_1}, \cdots, x_{i_n}) dP(x_1) \cdots dP(x_n),$$

f_0 满足 (1.3), 因而 $f_0(x)$ 即为 $f(X)$ 在给定 $T(x)$ 时的条件期望.

由上述定义容易验证, 对于期望的每一条通常的性质, 都有条件期望的一条性质与之相应.

引理 1.5 *设 T 是 $(\mathcal{X}, \mathcal{B})$ 到 $(\mathcal{T}, \mathcal{G})$ 的可测变换, $f, g, h, \cdots$ 均为 $(\mathcal{B}, P)$-可积的函数, $a, b, \cdots$ 为实数, 则有*

(i) $\mathrm{E}[af(X) + bg(X)|T] = a\mathrm{E}[f(X)|T] + b\mathrm{E}[g(X)|T]$ a.s.;

(ii) *若 f 非负*, 则 $\mathrm{E}[f(X)|T] \geqslant 0$ a.s.;

更一般地, 若 $f \leqslant g$, 则 $\mathrm{E}[f(X)|T] \leqslant \mathrm{E}[g(X)|T]$ a.s.;

(iii) $\mathrm{E}[h(T)f(X)|T] = h(T)\mathrm{E}[f(X)|T]$ a.s.;

(iv) $\mathrm{E}[\mathrm{E}(f(X)|T)] = \mathrm{E}[f(X)]$ (*条件期望的平滑性*);

(v) *若又设 S 是 $(\mathcal{T}, \mathcal{G})$ 到 $(\mathcal{S}, \mathcal{H})$ 的可测变换*, 则 $\mathrm{E}[\mathrm{E}(f(X)|T)|S] = \mathrm{E}[f(X)|S]$ a.s.;

(vi) *设 $0 \leqslant f_n \uparrow f$*, 则 $\mathrm{E}[f_n(X)|T] \uparrow \mathrm{E}[f(X)|T]$ a.s.;

(vii) *设所有 f_n 非负*, 则 $\mathrm{E}[\liminf_{n\to\infty} f_n(X)|T] \leqslant \liminf_{n\to\infty} \mathrm{E}[f_n(X)|T]$ a.s.;

(viii) *设 $|f_n| \leqslant h$*, $f_n \to f$ a.s., 则 $\mathrm{E}[f_n(X)|T] \to \mathrm{E}[f(X)|T]$ a.s. *成立*.

事实上, 性质 (i)—(v) 可直接由条件期望定义得出, 性质 (vi)—(viii) 可由积分的有关定理得出. 以 (vi) 为例. 记 $g_n(t) = \mathrm{E}[f_n(X)|t]$. 由性质 (ii), 序列 g_n 为 a.e. P^* 上升且有一 $\mathcal{G}$-可测的极限函数 g. 依条件期望的定义, 对 $\forall G \in \mathcal{G}$, $\int_G g_n(t) dP^*(t) = \int_{T^{-1}(G)} f_n(x) dP(x)$. 因 $g_n \uparrow g$ 而 $f_n \uparrow f$, 由单调收敛定理, 对 $\forall G \in \mathcal{G}$, $\int_G g(t) dP^*(t) = \int_{T^{-1}(G)} f(x) dP(x)$, 即 $g(T) = \mathrm{E}[f(X)|T]$, 因而 $\mathrm{E}[f_n(X)|T] \uparrow \mathrm{E}[f(X)|T]$ a.s..

在 (iii) 中对 $h(T)f(X)$ 取条件期望时, 因 T 是给定的, $h(T)$ 可作为一个常数提出来. 性质 (iv) 和 (v) 显示的是条件期望的平滑性质, 表明一次平均可以分两步进行, 在某些情况下这是一种方便的方法.

1.1.4 条件概率的定义和性质

以 I_B 记集合 B 的示性函数, 则 $P(X\in B)=\mathrm{E}[I_B(X)]$, 因而有如下定义.

定义 1.2 设 $B\in\mathcal{B}$, 则可以把在 T 给定的条件下 B 的**条件概率**定义为

$$P(B|T)=\mathrm{E}[I_B(X)|T]. \tag{1.5}$$

依据 (1.4), $P(B|t)$ 是一 $\mathcal{G}$-可测函数, 且满足方程

$$P[B\cap T^{-1}(G)]=\int_G P(B|t)dP^*(t),\quad \text{对所有}\ \ G\in\mathcal{G}. \tag{1.6}$$

如果不计它在一个 P^*-零测集上的取值, 则它是唯一确定的.

记 $\mathcal{B}_0=T^{-1}(\mathcal{G})$, 则可等价地定义 $P\big(B|T(x)\big)$ 为一 $\mathcal{B}_0$-可测函数, 且满足方程

$$P[B\cap B_0]=\int_{B_0} P\big(B|T(x)\big)dP(x),\quad \text{对所有}\ \ B_0\in\mathcal{B}_0. \tag{1.7}$$

关于条件概率的性质, 我们有如下引理.

引理 1.6 *设 T 是 $(\mathcal{X},\mathcal{B})$ 到 $(\mathcal{T},\mathcal{G})$ 的可测变换, $B_1,B_2,\cdots$ 为 $\mathcal{B}$ 中的集合, 则以下性质* a.e. $(\mathcal{G},P^*)$ *成立.*

(i) *若* $B_1\subset B_2$, *则* $P(B_1|t)\leqslant P(B_2|t)$;

(ii) *若* $B_1,B_2,\cdots$ *互斥, 则有*

$$P\Big(\bigcup_{i=1}^{\infty}B_i|t\Big)=\sum_{i=1}^{\infty}P(B_i|t);$$

若 $B_n\uparrow B$ *或* $B_n\downarrow B$, *则* $P(B_n|t)\ \to\ P(B|t)$;

(iii) $P(\mathcal{X}|t)=1,\quad P(\emptyset|t)=0.$

1.1.5 条件概率分布

依定义 1.2, 条件概率 $P(B|t)$ 应看作 B 固定时 t 的一个 $\mathcal{G}$-可测函数. 而在初等情形, 我们是将 t 视为固定的, 而将 $P(B|t)$ 看作 B 变化时在 $\mathcal{B}$ 上定义的一个集合函数. 引理 1.6 暗示了对固定的 t, 将 $P(B|t)$ 解释成 $\mathcal{B}$ 上的概率测度或概率分布的可能性. 但是, 为使上述性质成立, 对于每一组 $\{B_i\}$, 都要去除 t 的一个 P^*-零测集, 而这个例外集会随着 $\{B_i\}$ 的变化而变化, 它们的并可能已不再是一个 P^*-零测集. 幸而在 $(\mathcal{X},\mathcal{B})$ 为欧氏样本空间这个重要情形, 我们可以利用条件概率函数在那些例外集上的非唯一性, 适当选取条件概率, 使得引理 1.6 三条性质中 "a.e. $(\mathcal{G},P^*)$" 的尾巴可以去掉. 这就引出了下述定义.

定义 1.3 设 P 为样本空间 $(\mathcal{X},\mathcal{B})$ 上的一个概率分布, T 为取值于 $(\mathcal{T},\mathcal{G})$ 的一个统计量. 函数 $P(B,t),B\in\mathcal{B},t\in\mathcal{T}$ 称为一个**条件概率分布**, 如果它满足以下两个条件:

(i) 对每个固定的 $t\in\mathcal{T}$, $P(\cdot,t)$ 是 $\mathcal{B}$ 上的概率测度;

(ii) 对每个固定的 $B\in\mathcal{B}$, $P(B,t)$ 为在 $T=t$ 给定的条件下 B 的条件概率.

定理 1.7 设样本空间 $(\mathcal{X},\mathcal{B})$ 为欧氏的, 则满足定义 1.3 要求的条件概率分布 $P(B,t), B\in\mathcal{B}, t\in\mathcal{T}$ 存在.

证明 设 $\mathcal{X}$ 为 $\mathbf{R}^k$ 的一个Borel 子集. 令 $\tilde{P}(\cdot)=P(\cdot\cap\mathcal{X})$, 则可将概率测度 P 扩充到 $\mathbf{R}^k$ 的所有Borel 子集构成的 σ-域 $\mathcal{B}_k$. 故不失一般性, 可设 $(\mathcal{X},\mathcal{B})=(\mathbf{R}^k,\mathcal{B}_k)$. 为了简化, 我们将仅对一维情形给出证明.

对每个实数 x, 令 $F(x,t):=P\{(-\infty,x]|t\}$ 为任一选定的条件概率. 设 $\mathbf{Q}=\{r_1,r_2,\cdots\}$ 为有理数集, 则 $r_i<r_j\Rightarrow$ 存在一个 $\mathcal{G}$-可测的 P^* 零概集 (简称零概集) N_{ij}, 使得对于 N_{ij}之外的 t, 有 $F(r_i,t)\leqslant F(r_j,t)$, 因而对零概集 $N'=\bigcup N_{ij}$ 之外所有的 t, $F(x,t)$ 在有理数集上关于 x 是非减的. 类似地, 由引理 1.6(ii), 对在某个零概集 N'' 之外所有的 t, 有 $\lim_{n\to\infty}F(r_i+n^{-1},t)=F(r_i,t),\ i=1,2,\cdots$, $\lim_{n\to\infty}F(n,t)=1,\ \lim_{n\to\infty}F(-n,t)=0$. 因此, 对零概集 $N=N'\cup N''$ 之外所有的 t, $F(\cdot,t)$ 在有理数集上满足分布函数的所有性质. 易见, 对 N 之外的 t, $F(\cdot,t)$ 可以唯一地扩充为 $\mathbf{R}$ 上右连续的分布函数 $F_*(\cdot,t)$, 后者确定了 $\mathcal{B}$ 上的一个概率测度 $P_*(\cdot,t)$. 对 $t\in N$, 可选取 $\mathcal{B}$ 上的任一概率测度作为 $P_*(\cdot,t)$.

为了证明定理, 只需证明对每个固定的 $B\in\mathcal{B}$, $P_*(B,t)$ 为在给定 t 的条件下 B 的条件概率, 即 $P_*(B,t)$ 满足

$$P_*(B,t)=P(B|t),\quad \text{a.e. } (\mathcal{G},P^*).$$

设使得上式成立的 B 组成集类 $\mathcal{M}$. 由 P_* 的定义, 任一有理区间 $(a,b]$ 属于 $\mathcal{M}$. 由 P_* 为测度, 及引理 1.6(ii), $\mathcal{M}$ 包含由有理区间的所有有限并构成的域 $\mathcal{F}$. 最后, $\mathcal{M}$ 为一单调类, 因而包含由 $\mathcal{F}$ 生成的 σ-域 $\mathcal{B}$. 至此, 条件分布的构造即告完成 (参看Lehmann (1986), 第 49 页, 或陈希孺 (1981, 1997), 第 36—38 页). ■

关于条件分布 $P(B,t)$ 与条件期望之间的联系, 我们有如下定理.

定理 1.8 设 $P(B,t):\mathcal{B}\times\mathcal{T}\mapsto[0,1]$ 为一条件分布, $\mathrm{E}|f(X)|<\infty$, 则

$$\mathrm{E}[f(X)|t]=\int_{\mathcal{X}}f(x)P(dx,t),\ \text{a. e. } (\mathcal{G},P^*).$$

证明很简单. 若 $f(x)=I_B(x), B\in\mathcal{B}$, 则上式成立. 由引理 1.5, 它对简单函数, 继而对任何可积函数都成立.

定理 1.9 设统计量 T 的定义域和值域空间 $(\mathcal{X},\mathcal{B})$ 和 $(\mathcal{T},\mathcal{G})$ 均为欧氏的, 则存在条件概率分布的一种选取 $P(B,t)$ 以及一个 P^*-零测集 N, 当 $t\notin N$ 时, 有 $P(B^{(t)},t)=1$, 其中 $B^{(t)}=T^{-1}(\{t\})=\{x:T(x)=t\}$.

这一性质的直观意义很清楚, 给定了 $T(x)=t$, 等于说限制了 x 必须在 $B^{(t)}$ 中, 因而后者的条件概率应为 1.

证明 设对每一个 t, $P(B,t)$ 是 $\mathcal{B}$ 上的一个概率分布, 它的存在性由定理 1.7 保证. 给定 $G\in\mathcal{G}$, 示性函数 I_G 是 $\mathcal{G}$-可测的, 且

$$\int_{G'}I_G(t)P^*(dt)=P^*(G'\cap G)=P(T^{-1}G'\cap T^{-1}G),\quad\forall\, G'\in\mathcal{G}.$$

由条件概率的定义, 除去一个 P^*-零测集 $N(G)$, 当 $t \notin N(G)$ 时, $I_G(t) = P(T^{-1}G, t)$. 用 $\mathcal{H}$ 记 $\mathcal{T}$ 中所有有理区间组成的可列类, 则存在一个 P^*-零测集 $N = \bigcup_{G\in\mathcal{H}} N(G)$, 当 $t \notin N$ 时, 对所有有理区间 G,

$$I_G(t) = P(T^{-1}G, t).$$

两个集合函数 $P(T^{-1}G, t)$ 和 $I_G(t)$ 都是 σ-域 $\mathcal{G}$ 上的概率分布, 当 $t \notin N$ 时, 两者在所有有理区间上一致, 因而在 $\mathcal{G}$ 上一致. 特别地, 对于 $t \notin N$, 由单点集 $\{t\} \in \mathcal{G}$, 知 $P(B^{(t)}, t) := P(T^{-1}(G), t) = I_{\{t\}}(t) = 1$. ■

定理 1.10 *采用定理 1.9 选取的条件分布和 P^*-零测集 N, 则当 $t \notin N$ 时, 对任何 P-可积函数 $f(X)$,*

$$\mathrm{E}[f(X)|t] = \int_{B^{(t)}} f(x)P(dx, t). \tag{1.8}$$

特别地, 当 $t \notin N$ 且 h 为 $\mathcal{G}$ 可测函数时,

$$\mathrm{E}[h(T)f(X)|t] = h(t)\mathrm{E}[f(X)|t]. \tag{1.9}$$

此处, 定理 1.8 和引理 1.5(iii) 的结论得到了加强: 使等式不成立的零测度的例外集可以取得与函数 f 及 h 无关.

作为定理的一个推论, 对所有的 $t \notin N$, $\mathrm{E}[h(T)|t] = h(t)$, 特别地, $P(T \in G|t) = 1$ 或 0, 视 $t \in G$ 或 $t \notin G$ 而定.

以下分别用 P^X 和 P^T 记 X 和 T 的分布. 用 $P^{X|t}$ 记 $\mathcal{B}$ 上由定理 1.7 保证的条件分布 $P(\cdot, t)$ 的任一代表.

条件分布 $P^{X|t}$ 仍然与初等情形不同, 因为它是定义在 $(\mathcal{X}, \mathcal{B})$ 上, 而不是定义在集合 $B^{(t)}$ 及其 Borel 子集的 σ-域 $\mathcal{B}^{(t)}$ 上. 但在定理 1.9 的条件下, 对 $t \notin N$, $P^{X|t}(B^{(t)}) = 1$ 蕴含了

$$P^{X|t}(B) = P^{X|t}(B \cap B^{(t)}), \quad B \in \mathcal{B}.$$

因此 $P^{X|t}$ 就可用分布 $\bar{P}^{X|t}$ 代替, 后者定义在 $(B^{(t)}, \mathcal{B}^{(t)})$ 上, 且对 $B^{(t)}$ 的子集, 它与 $P^{X|t}$ 赋予了它们相同的概率. 由条件期望的平滑性 (引理 1.5(iv)),

$$\mathrm{E}[f(X)] = \int_{\mathcal{T}-N} \Big\{ \int_{B^{(t)}} f(x)d\bar{P}^{X|t}(x) \Big\} dP^T(t) \tag{1.10}$$

对所有可积函数 f 成立. 反之, 这一事实也可用来作为 $\bar{P}^{X|t}$ 的定义.

今设 $(\mathcal{X}, \mathcal{B})$ 为一乘积空间 $(\mathcal{T}\times\mathcal{Y}, \mathcal{G}\times\mathcal{C})$, 则 $B^{(t)} = \{t\}\times\mathcal{Y}$. 对 $t \notin N$, 条件分布 $\bar{P}^{X|t}$ 导出了 $(\mathcal{Y}, \mathcal{C})$ 上的一个分布, 与初等情形类似记成 $P^{Y|t}$. 在这种情况下, 定义也可扩充到整个 $\mathcal{T}$ 上, 只需对所有 $t \in N$, 让 $P^{Y|t}$ 赋予某一特定的公共点 y_0 以概率 1. 由这些定义, (1.10) 即为

$$\mathrm{E}[f(T, Y)] = \int_{\mathcal{T}} \Big\{ \int_{\mathcal{Y}} f(t, y)dP^{Y|t}(y) \Big\} dP^T(t). \tag{1.11}$$

作为一个应用,我们引进以下的引理, 这在指数族的分布理论中将会用到.

引理 1.11 *设 $(\mathcal{T}, \mathcal{G})$ 和 $(\mathcal{Y}, \mathcal{C})$ 均为欧氏空间, 命 $P_0^{(T,Y)}$ 是在乘积空间 $(\mathcal{X}, \mathcal{B}) =$*

$(\mathcal{T}\times\mathcal{Y},\mathcal{G}\times\mathcal{C})$ 上的一个分布. 假定 $(\mathcal{X},\mathcal{B})$ 上的另一分布为

$$dP_1(t,y)=a(y)b(t)dP_0(t,y),$$

此处, 对所有的 y, $a(y)>0$, 则对 P_1 而言, T 的边缘分布以及给定 t 时 Y 的条件分布由下式给出:

$$dP_1^T(t)=b(t)\Big\{\int_{\mathcal{Y}}a(y)dP_0^{Y|t}(y)\Big\}dP_0^T(t),$$

$$dP_1^{Y|t}(y)=\frac{a(y)dP_0^{Y|t}(y)}{\int_{\mathcal{Y}}a(y')dP_0^{Y|t}(y')}.$$

证明 引理的第一个结论由下式得出:

$$P_1(T\in G)=\mathrm{E}_1[I_G(T)]=\mathrm{E}_0[I_G(T)a(Y)b(T)]=\int_G b(t)\Big\{\int_{\mathcal{Y}}a(y)dP_0^{Y|t}(y)\Big\}dP_0^T(t).$$

为验证第二个结论, 只需证明对任何可积函数 f, 其期望值 $\mathrm{E}_1[f(T,Y)]$ 满足 (1.11), 而这是很容易做到的. ■

如上所述, 在条件期望的定义中没有提供构造方法, 因而涉及条件期望、条件概率和条件分布的计算往往很难, 需要借助于直观和启发式的推导技巧去求解, 必要时再用定义加以验证, 无一定之规可循. 今举例如下.

例 1.2 设 $X_1,\cdots,X_n$ i.i.d., X_1 有指数分布 $\mathrm{e}^{-x}I(x>0)dx$, 统计量 $X_{(1)}=\min\{X_1,\cdots,X_n\}$. 试求 $X_1-X_{(1)}$ 在 $X_{(1)}$ 给定时的条件分布.

首先, $P(X_1-X_{(1)}=0|X_{(1)})=1/n$.

为求上述条件分布, 只需对任一正数 u, 算出 $P(X_1-X_{(1)}>u|X_{(1)})$. 由对称性,

$$\begin{aligned}P(X_1-X_{(1)}>u|\ X_{(1)})&=\frac{1}{n}\sum_{i=1}^{n}P(X_i-X_{(1)}>u|X_{(1)})\\&=\frac{1}{n}\sum_{i=1}^{n}P(X_{(i)}-X_{(1)}>u|X_{(1)}).\end{aligned}$$

当 $X_{(1)}$ 给定时, $(X_{(i)}-X_{(1)},i=2,\cdots,n)$ 的条件分布, 与从指数分布 $e^{-x}I(x>0)dx$ 中取出的 i.i.d. 样本 $Y_1,\cdots,Y_{n-1}$ 的次序统计量的分布一致. 故

$$P(X_1-X_{(1)}>u|X_{(1)})=\frac{1}{n}\sum_{i=1}^{n-1}P(Y_i>u)=\frac{n-1}{n}\mathrm{e}^{-u}.$$

读者不难写出条件分布函数. 需要特别注意的是, 该函数在原点有一个跳跃, 其跃度为 $1/n$, 因而 $P(X_1-X_{(1)}>0|X_{(1)})=(n-1)/n$ 而不是 1.

1.2 样本空间与分布族

数理统计学的任务, 是研究如何有效地收集、整理和分析带有随机性的数据, 以对所考察的问题作出推断或预测, 直至为采取一定的决策和行动提供依据和建议(参看《中国大百科全书》数学卷). 本书只讨论在已有数据即样本的情况下如何进行统

计分析的问题. 至于如何通过观测或试验收集数据, 则是数理统计学的两个重要分支, 即试验设计和抽样调查的研究内容.

1.2.1 样本空间与样本分布族

简言之, 有关的统计数据源于一个**随机元** X. 从数学上考虑, 首先要有一个可测空间 $(\mathcal{X},\mathcal{B})$, 称之为**样本空间**, 其中集合 $\mathcal{X}$ 包含了随机元 X 可能取值的范围, $\mathcal{B}$ 是 $\mathcal{X}$ 的某些子集构成的一个 σ-域, 这样的样本空间就是随机元 X 及其概率分布 P 的居所. $\mathcal{B}$ 中的集合称为 $\mathcal{B}$-可测集, 简称为可测集. 依 X 的概率分布从 $\mathcal{X}$ 中随机抽取的一个元素 x, 相应于随机元 X 的一个观测值, 称之为一个**样本值**. 在许多常见的统计问题中, X 为一有限维 (例如 n 维) 随机向量, 此时可取 n 维欧氏空间 $\mathbf{R}^n$ 或它的一个适当的 Borel 子集作为 $\mathcal{X}$, 而将其所有 Borel 子集构成的 σ-域 $\mathcal{B}_n$ 取作 $\mathcal{B}$, 称这样的样本空间为欧氏的. 在统计问题中, 随机元 X 的概率分布 P 总是未知或部分未知的. 它的某些未知方面正是我们需要加以研究的. 因此我们可以假定 P 属于某个分布族 $\mathcal{P}$. $\mathcal{P}$ 在具体问题中可以有确切的界定, 它包含的信息表述了我们在理论或经验方面对该问题的知识.

在数理统计中, 通常把随机元 X 称为总体, 其分布 P 自然被称为**总体分布**, $\mathcal{P}$ 则称为**总体分布族**. 另一方面, 样本 x 又是依 X 的概率分布从 $\mathcal{X}$ 中抽取的, 因此在此一般架构下, P 和 $\mathcal{P}$ 也被称为**样本分布**和**样本分布族**. 但在一种重要场合这两个概念可以有不同的理解. 例如, 假定 $X_1,\cdots,X_n$ 是一组 i.i.d. 观测数据, 每个变量都服从分布 $Q\in\mathcal{Q}$. 用 Q^n 记 Q 的 n 重乘积测度. 一方面, 我们可以把 n 维随机向量 $X:=(X_1,\cdots,X_n)$ 当作一个随机元. 此时若设 $x=(x_1,\cdots,x_n)$ 为 X 的一个样本, 则总体分布族和样本分布族一致, 同为 $\{X\sim Q^n:\ Q\in\mathcal{Q}\}$. 另一方面, 若我们把服从分布 Q 的一个变量 $\widetilde{X}$ 当作一个随机元, 把 $x_1,\cdots,x_n$ 看成从总体中 n 次独立抽样所得数据组成的一个 (合) 样本, 则在这一特殊场合, 总体分布与样本分布有不同的含义. 这样的一个样本称为取自总体 $\widetilde{X}$ 的一个**独立随机样本**或**简单随机样本**, 简记为 i.i.d. 样本, n 称为样本量.

不难看出, 样本具有两重性. 一方面, 样本是基于总体分布按一定方式抽取的, 因而为一随机变量; 但在特定的一次抽样中所取的是一组具体的数值, 称为样本值.

统计中特别感兴趣的分布族是所谓的**受控分布族**. 设 $\mathcal{P}$ 为定义于可测空间 $(\mathcal{X},\mathcal{B})$ 上的一个分布族, μ 为此空间上的一个 σ-有限测度, 如果对任何 $C\in\mathcal{B}$, $\mu(C)=0\Rightarrow P(C)=0$, 对任何 $P\in\mathcal{P}$, 则称 $\mathcal{P}$ 受控于 μ, 记为 $\mathcal{P}\ll\mu$. 依 Radon-Nikodym 定理, 对每个 $P\in\mathcal{P}$, 存在一个定义于 $\mathcal{X}$ 上的非负 $\mathcal{B}$-可测函数 f_P, 使 $dP=f_P d\mu$, 称 f_P 为 P 对测度 μ 的密度.

在欧氏样本空间的情况下, 若 X 为离散随机向量, 一般取 μ 为计数测度, 而在连续随机向量情形, 则取 μ 为有关欧氏空间上的 Lebesgue 测度. 一个重要的情形是, 在分布族 $\{dP(x)=f_P(x)d\mu(x),P\in\mathcal{P}\}$ 中, $S(f_P):=\{x:x\in\mathcal{X},f_P(x)>0\}$ 与

$P \in \mathcal{P}$ 无关, **此时称分布族 $\mathcal{P}$ 有共同支撑**.

样本空间连同在它上面定义的样本分布族, 构成了一个统计问题的基本要素. 这些要素的指定, 即为有关问题提供了一个确定的统计模型. 模型的概率背景派生出一个事实, 即现代统计分析从理论和方法的建立, 直到所得结果的解释, 都离不开概率论的理论支撑.

关于样本分布族 $\mathcal{P}$, 按其复杂程度和有关统计问题的性质, 可以分成以下几类.

参数族 当 $\mathcal{P}$ 中的分布的一般数学形式已知, 但包含若干个未知参数时, 称 $\mathcal{P}$ 为参数族. 形式上说, 即存在一个 r-维欧氏空间 $\mathbf{R}^r$ 的 Borel 子集 Θ, 使得 Θ 与 $\mathcal{P}$ 之间能建立一一对应. 可将与 θ 对应的分布记为 P_θ, 而将分布族 $\mathcal{P}$ 记为 $\{P_\theta, \theta \in \Theta\}$, θ 称为**参数**, 而 Θ 称为**参数空间**. 此处的 "一一对应" 是对参数表示的可识别性要求, 也就是要求 $\theta_1 \neq \theta_2 \Rightarrow P_{\theta_1} \neq P_{\theta_2}$. 为方便计, 也可以把包含 Θ 的一个 r-维 Borel 子集取为参数空间, 通常维数 r 要取成最小的, 此 r 即称为所论统计模型的维数. 这个维数当然不能与分布 P 的维数混为一谈.

对受控的参数族, 可采用相应的记号, 例如 $\{f(x,\theta)d\mu(x),\ \theta \in \Theta\}$.

初等统计中讨论过的一些常见分布族, 诸如二项分布族、Poisson 分布族、正态分布族、指数分布族和均匀分布族等都是参数族.

非参数族 当 $\mathcal{P}$ 中的分布不能通过有限个未知参数去刻画时, $\mathcal{P}$ 称为非参数分布族. 例如, 所有一维对称分布组成的分布族, 分布函数连续的所有一维分布组成的分布族, 都是非参数族的典型例子. 若一个统计模型涉及非参数样本分布族, 则有关的统计问题称为非参数的, 研究非参数统计问题的统计学分支称为**非参数统计**.

在现代统计的文献中, 研究对象早已不限于单纯的参数模型或单纯的非参数模型, 一些复杂的统计数据和统计模型引起了广泛的重视. 半参数统计模型就是一个例子. 在计量经济学等学科的研究中, 就产生了大量的**半参数统计模型**. 因而一些新的分支学科的分野, 也常常要考虑到统计模型的分类, 以及为了解决这些统计问题所使用的统计方法和数学工具.

本书涉及的分布族主要是参数族. 今后我们也常将 X 服从分布 P_θ 简记为 $X \sim P_\theta$.

1.2.2 指数型分布族

在初等统计中, 主要讨论与一些具体分布, 特别是与正态分布有关的统计问题. 其他像二项分布, Poisson 分布, 指数分布, Gamma 分布等, 也占有重要地位. 这些看起来不一样的分布族, 其实可以统一在一类更为广泛的指数型分布族中.

定义 1.4 欧氏样本空间 $(\mathcal{X}, \mathcal{B})$ 上的一个分布族 $\{P_\theta, \theta \in \Theta\}$ 称为**指数型分布族**, 简称 **指数族**, 如果它可以表示为

$$dP_\theta(x) = C(\theta)\exp\left\{\sum_{j=1}^{k} Q_j(\theta)T_j(x)\right\}h(x)d\mu(x), \tag{1.12}$$

其中, μ 为 $\mathcal{B}$ 上的一个 σ-有限测度, $T_1,\cdots,T_k$ 和 $h\geqslant 0$ 都是实值 Borel 可测函数, $Q_1,\cdots,Q_k$ 和 $C>0$ 都是参数空间 Θ 上的实值函数.

上述定义隐含了以下条件成立:

$$0<1/C(\theta)=\int_{\mathcal{X}}\exp\left\{\sum_{j=1}^{k}Q_j(\theta)T_j(x)\right\}h(x)d\mu(x)<\infty,\quad \theta\in\Theta.$$

参数空间最大可以拓展到由满足上式的所有 θ 组成的集合上. 易见, 指数族是具有共同支撑的分布族.

为了将作为寿命分布的一维指数分布族与指数族加以区别, 今后将形如

$$dP_{(\alpha,\beta)}(x)=\beta^{-1}\exp\{-(x-\alpha)/\beta\}I_{(\alpha,\infty)}(x)dx,\quad \alpha\in(-\infty,\infty),\beta>0 \tag{1.13}$$

的分布族记为指数分布族 $\{E(\alpha,\beta)\}$.

易见, 本小节开头提到的常见分布, 连同从中取出的样本 $X=(X_1,\cdots,X_n)$ 的分布, 都属于指数族. 例如, 设 $X_1,\cdots,X_n$ 为 i.i.d., $X_1\sim N(\alpha,\sigma^2),-\infty<\alpha<\infty,0<\sigma^2<\infty$, 则 X 关于 n-维 Lebesgue 测度的密度函数可以表示为

$$f(x|\alpha,\sigma^2)=C(\alpha,\sigma^2)\exp\left\{-\frac{1}{2\sigma^2}\sum_{j=1}^{n}x_j^2+\frac{\alpha}{\sigma^2}\sum_{j=1}^{n}x_j\right\}, \tag{1.14}$$

其中 $C(\alpha,\sigma^2)=(2\pi\sigma^2)^{-n/2}\exp\{-n\alpha^2/(2\sigma^2)\}$.

若用 μ 记整数集上的计数测度, 则在二项分布族的场合, X 的分布为

$$dP_p(x)=\binom{n}{x}p^x(1-p)^{n-x}d\mu(x)=(1-p)^n\exp\left\{x\log\frac{p}{1-p}\right\}\binom{n}{x}d\mu(x). \tag{1.15}$$

参数空间为 $(0,1)$. 而在 Poisson 分布族的情形,

$$X\sim dP_\lambda(x)=\mathrm{e}^{-\lambda}(\lambda^x/x!)d\mu(x)=\mathrm{e}^{-\lambda}\cdot\exp\{x\log\lambda\}\cdot(1/x!)d\mu(x). \tag{1.16}$$

参数空间为 $(0,\infty)$.

采用更加自然的参数并将 $h(x)$ 吸收进 $d\mu(x)$ 之中, 则可以把指数族写成

$$dP_\theta(x)=C(\theta)\exp\left\{\sum_{j=1}^{k}\theta_jT_j(x)\right\}d\mu(x)=C(\theta)\exp\left\{\theta^\tau T(x)\right\}d\mu(x) \tag{1.17}$$

的形式, 其中 $\theta=(\theta_1,\cdots,\theta_k)^\tau$, $T(x)=\big(T_1(x),\cdots,T_k(x)\big)^\tau$. 今后, A^τ 总是用来记一个矩阵 A 的转置. 在指数族的理论研究中, 一般取 (1.17) 作为出发点, 称之为指数族的自然形式. 此时把满足

$$\int_{\mathcal{X}}\exp\left\{\sum_{i=1}^{k}\theta_iT_i(x)\right\}d\mu(x)<\infty \tag{1.18}$$

的所有 θ 组成的集合 Θ 称为**自然参数空间**.

为了今后的需要, 我们现在考虑指数族的某些性质. 由 Hölder 不等式立即得到如下引理.

引理 1.12 *一个指数族的自然参数空间 Θ 是 $\mathbf{R}^k$ 空间中的一个凸集.*

如果 Θ 落在一个低于 k-维的线性空间内, 则 (1.17) 可以改写成 T 仅包含少于 k 个分量的形式. 因此, 不失一般性, 可以假定 Θ 即为 k-维的. 为方便计, 用 B° 记集合 $B(\subset \mathbf{R}^k)$ 在 $\mathbf{R}^k$ 中所有内点的集合, 即其内部. 下面的定理揭示了指数族良好的分析性质.

定理 1.13 设 $a=(a_1,\cdots,a_k)^\tau$ 为指数族 (1.17) 的自然参数空间 Θ 在 $\mathbf{R}^k$ 中的一个内点, 则函数

$$H(\theta)=1/C(\theta)=\int_{\mathcal{X}}\exp\left\{\sum_{j=1}^{k}\theta_jT_j(x)\right\}d\mu(x) \tag{1.19}$$

(a) 在 a 点连续;

(b) 它在 a 点的各阶偏导数都可在积分号下去取;

(c) 若将上述积分看成复变数 $\theta_j=\xi_j+\mathrm{i}\eta_j$ $(j=1,\cdots,k)$ 的函数, 则它在复区域

$$U=\{\theta=(\xi_1+\mathrm{i}\eta_1,\cdots,\xi_k+\mathrm{i}\eta_k)^\tau:(\xi_1,\cdots,\xi_k)^\tau\in\Theta^{\circ}\}$$

上是每个变元的解析函数. 此处记 $\mathrm{i}=\sqrt{-1}$.

证明 (a) 因为 $a\in\Theta^{\circ}$, 故存在一 k-维闭立方体 $A\subset\Theta^{\circ}$, 使 $a\in A^{\circ}$. 每个 $\theta\in A^{\circ}$ 都可表示为 A 的 $m=2^k$ 个顶点 v_j 的凸组合 $\sum_{j=1}^{m}\alpha_jv_j$ 的形式, 其中每个 α_j 非负, $\sum\alpha_j=1$. 由凸函数的 Jensen 不等式,

$$\exp\{\theta^\tau T(x)\}\leqslant\sum_{j=1}^{m}\alpha_j\exp\{v_j^\tau T(x)\}\leqslant\sum_{j=1}^{m}\exp\{v_j^\tau T(x)\}:=G(x).$$

由 A 的取法, $\int G(x)d\mu(x)<\infty$, 故由控制收敛定理, 对趋于 a 的任何点列 θ_n,

$$\lim_{\theta_n\to a}\int\exp\{\theta_n^\tau T(x)\}d\mu(x)=\int\exp\{a^\tau T(x)\}d\mu(x).$$

这就证明了 (a).

为证 (b), 考虑 H 对 θ_1 的一阶偏导数. 记 $d\mu_1(x):=\exp\big\{\sum_{j=2}^{k}\theta_jT_j(x)\big\}d\mu(x)$, 则问题简化为 $\theta=\theta_1$ 即 Θ 和 T 均为一维的情形. 设 $a\in\Theta^{\circ}$, 则存在 $c>0$ 使 $a\pm c\in\Theta^{\circ}$. 差商

$$\big\{H(a+h)-H(a)\big\}/h=\int_{\mathcal{X}}\mathrm{e}^{aT}\{(\mathrm{e}^{hT}-1)/h\}d\mu_1(x).$$

当 $|h|\leqslant c$ 时有 $|(\mathrm{e}^{hT}-1)/h|\leqslant\mathrm{e}^{c|T|}/c$, 故被积函数被

$$\exp\{aT+c|T|\}/c\leqslant[\exp\{(a+c)T\}+\exp\{(a-c)T\}]/c$$

所控制, 由 $a\pm c\in\Theta^{\circ}$, 知上述函数关于 $d\mu_1(x)$ 可积, 故由控制收敛定理, 对趋于 0 的任何点列 h_n, H 的差商序列趋于

$$\int T(x)\exp\{aT(x)\}d\mu_1(x).$$

这就对一阶偏导数情形证明了 (b), 对于高阶偏导数, 可利用归纳法类似地证明.

在复平面的一个开区域上定义的复变函数是解析的, 当且仅当它是可导的. 因而只需沿用上述证明方法并稍加改变, 即可建立 (c) 中的论断. 证毕. ■

注 1.1 若 θ 是自然参数空间在 $\mathbf{R}^k$ 中的内点, 则 T_j 在参数 θ 之下的各阶矩都存在有限, 且可通过在积分号下求导算出. 例如, 在 (1.19) 两边对 θ_1 求导, 得到

$$-C^{-2}(\theta)\partial C(\theta)/\partial\theta_1 = \int \exp\{\theta^\tau T(x)\}T_1(x)d\mu(x),$$

故有 $\mathrm{E}_\theta[T_1(X)] = -\partial \log C(\theta)/\partial\theta_1$, 其中 E_θ 表示期望是在 P_θ 之下取的. 类似可得 $\mathrm{Cov}_\theta[T_1(X), T_2(X)] = -\partial^2 \log C(\theta)/(\partial\theta_1\partial\theta_2)$.

注 1.2 采用向量和矩阵的记号, 若 $\theta \in \Theta^{\mathrm{o}}$, 则 $T(X)$ 在 P_θ 之下的协方差阵

$$\mathrm{Cov}_\theta[T(X)] = -\frac{\partial^2}{\partial\theta\partial\theta^\tau}\log C(\theta). \tag{1.20}$$

为了明确起见, 在涉及指数族 (1.17) 的问题中, 一般都要引进如下的假定:

(i) 假定 Θ 在 $\mathbf{R}^k$ 中有内点, 此时用 Θ^{o} 记所有这些内点的集合;

(ii) 假定对任何 $\theta \in \Theta^{\mathrm{o}}$, 不存在非零的常数向量 a, 使得线性组合 $a^\tau T(X)$ 在 P_θ 之下退化为一常数, 或者等价地, 即假定 $T(X)$ 在 P_θ 之下的协方差阵 $\mathrm{Cov}_\theta[T(X)]$ 为非奇异的. 此时, 在自然参数空间的内部, 函数 $\log C(\theta)$ 是严格凹的.

通常**称满足这样两个条件的指数族是满秩的**.

注 1.3 定理 1.13 容易推广到下述一般积分

$$G(\theta) = \int_{\mathcal{X}} g(x)\exp\{\theta^\tau T(x)\}d\mu(x), \tag{1.21}$$

其中, g 为定义于 $\mathcal{X}$ 上的 Borel 可测函数, $\Theta = \{\theta : \int |g(x)|\exp\{\theta^\tau T(x)\}d\mu(x) < \infty\}$, Θ^{o} 为 Θ 的内部, 而 $a \in \Theta^{\mathrm{o}}$.

关于指数族的边缘分布和条件分布, 有下面的引理. 这一引理在后面要用到.

引理 1.14 (i) *设 X 的分布为形如*

$$dP_\theta(x) = C(\theta)\exp\left\{\sum_{i=1}^k \theta_i T_i(x)\right\}d\mu(x)$$

的指数族. 记 μ^* 为测度 μ 在变换 $T = (T_1, \cdots, T_k)^\tau : (\mathcal{X}, \mathcal{B}) \mapsto (\mathbf{R}^k, \mathcal{B}_k)$ 下的导出测度, 则 T 的分布为形如

$$dP_\theta^T(t) = C(\theta)\exp\{\theta^\tau t\}d\mu^*(t)$$

的指数族. 此处, $\theta = (\theta_1, \cdots, \theta_k)^\tau, t = (t_1, \cdots, t_k)^\tau$.

(ii) *设 X 的分布为形如*

$$dP_{\theta,\vartheta}(x) = C(\theta,\vartheta)\exp\left\{\sum_{i=1}^r \theta_i U_i(x) + \sum_{j=1}^s \vartheta_j V_j(x)\right\}d\mu(x)$$

的指数族, 则分别存在 s-维和 r-维欧氏空间上的测度 λ_θ 和 ν_v, 使得

(a) *$V = (V_1, \cdots, V_s)^\tau$ 的分布为形如*

$$dP^V_{\theta,\vartheta}(v) = C(\theta,\vartheta)\exp\left\{\sum_{j=1}^{s}\vartheta_j v_j\right\}d\lambda_\theta(v) \tag{1.22}$$

的指数族.

(b) 给定 $V=v$ 时 $U=(U_1,\cdots,U_r)^\tau$ 的条件分布为形如

$$dP^{U|v}_{\theta}(u) = C_v(\theta)\exp\left\{\sum_{i=1}^{r}\theta_i u_i\right\}d\nu_v(u) \tag{1.23}$$

的指数族. 它不依赖于 ϑ.

证明 (i) 任取 $B\in\mathcal{B}_k$, 则由引理 1.2 中的积分变换公式,

$$P_\theta\{T(X)\in B\} = C(\theta)\int_{T^{-1}(B)}\exp\{\theta^\tau T(x)\}d\mu(x) = C(\theta)\int_B\exp\{\theta^\tau t\}d\mu^*(t),$$

即 (i) 中的结论成立.

(ii) 设 μ^* 为 μ 在变换 $(U,V):(\mathcal{X},\mathcal{B})\mapsto(\mathbf{R}^{r+s},\mathcal{B}_{r+s})$ 下的导出测度, 由 (i), (U,V) 的分布为

$$dP_{\theta,\vartheta}(u,v) = C(\theta,\vartheta)\exp\{\theta^\tau u+\vartheta^\tau v\}d\mu^*(u,v),$$

此处 $\theta=(\theta_1,\cdots,\theta_r)^\tau,\vartheta=(\vartheta_1,\cdots,\vartheta_s)^\tau$, 对于 u 和 v 也有类似的向量记号.

设 (θ_*,ϑ_*) 为自然参数空间中的一点, 则

$$dP_{\theta,\vartheta}(u,v) = [C(\theta,\vartheta)/C(\theta_*,\vartheta_*)]\exp\{(\theta-\theta_*)^\tau u+(\vartheta-\vartheta_*)^\tau v\}dP_{\theta_*,\vartheta_*}(u,v).$$

令

$$d\lambda_\theta(v) = [1/C(\theta_*,\vartheta_*)]\exp(-\vartheta_*^\tau v)\Big\{\int\exp[(\theta-\theta_*)^\tau u]dP^{U|v}_{\theta_*,\vartheta_*}(u)\Big\}dP^V_{\theta_*,\vartheta_*}(v),$$

$$d\nu_v(u) = \exp\{-\theta_*^\tau u\}dP^{U|v}_{\theta_*,\vartheta_*}(u),$$

由引理 1.11 即可推得 (ii) 中的结论. ■

1.2.3 若干常用分布族

为了今后使用的方便, 本小节简要介绍若干常用分布族及其记号.

(1) 二项分布 $B(n,p),0\leqslant p\leqslant 1$:

$$P(X=x) = \binom{n}{x}p^x(1-p)^{n-x},\quad x=0,1,\cdots,n.$$

当 $n=1$ 时, 即得两点分布, 又称 Bernoulli 分布:

$$P(X=x) = p^x(1-p)^{1-x},\quad x=0,1.$$

(2) Poisson 分布 $P(\lambda),\lambda>0$:

$$P(X=x) = \lambda^x\mathrm{e}^{-\lambda}/x!,\quad x=0,1,\cdots.$$

(3) 负二项分布 $NB(r,p),0<p<1$. 可以有两种定义. 设在成功概率为 p 的重复的 Bernoulli 试验中, 试验一直进行到第 r 次成功为止. 若用 X 记试验的总次数, 则

$$P(X=x) = \binom{x-1}{r-1}p^r(1-p)^{x-r},\quad x=r,r+1,\cdots.$$

若用 Y 表示到第 r 次成功为止失败的次数, 则

$$P(Y=y)=\binom{y+r-1}{r-1}p^r(1-p)^y,\quad y=0,1,\cdots.$$

按照第二种定义, 当 $r=1$ 时即为几何分布.

(4) 多项分布 $M(n;p_1,\cdots,p_r),\forall\ p_j>0,\ \sum_{j=1}^r p_j=1$:

$$P(X_1=x_1,\cdots,X_r=x_r)=\frac{n!}{x_1!\cdots x_r!}p_1^{x_1}\cdots p_r^{x_r},$$

此处, 每个 x_j 为非负整数, 且 $\sum_{j=1}^r x_j=n$.

(5) 正态分布 $N(\alpha,\sigma^2),-\infty<\alpha<\infty,\ \sigma^2>0$: 密度函数为

$$f(x;\alpha,\sigma^2)=(2\pi\sigma^2)^{-1/2}\exp\{-(x-\alpha)^2/(2\sigma^2)\},\quad -\infty<x<\infty.$$

此处的密度是对 Lebesgue 测度而言的, 下面不再赘述.

标准正态记为 $N(0,1)$, 其分布函数记为 Φ.

多元正态分布 $N(\alpha,\Sigma),\alpha\in\mathbf{R}^n,\Sigma:n\times n,\Sigma>0$: 密度函数为

$$f(x;\alpha,\Sigma)=(2\pi)^{-n/2}\big|\Sigma\big|^{-1/2}\exp\{-(x-\alpha)^\tau\Sigma^{-1}(x-\alpha)/2\},\quad x\in\mathbf{R}^n,$$

此处 $\big|\Sigma\big|=\det\Sigma$. 其特征函数为 $\psi(t)=\exp\{\mathrm{i}\alpha^\tau t-t^\tau\Sigma t/2\}$.

其中二元正态 $N(\alpha,\beta;\sigma_1^2,\sigma_2^2,\rho)$ 的密度函数为

$$f(x;\theta)=\frac{1}{2\pi\sigma_1\sigma_2\sqrt{1-\rho^2}}\exp\Big\{-\frac{1}{1-\rho^2}\Big[\frac{(x-\alpha)^2}{2\sigma_1^2}+\frac{(y-\beta)^2}{2\sigma_2^2}-\frac{\rho(x-\alpha)(y-\beta)}{\sigma_1\sigma_2}\Big]\Big\},$$

此处, $\theta=(\alpha,\beta;\sigma_1^2,\sigma_2^2,\rho):\ \alpha\in\mathbf{R},\beta\in\mathbf{R},\sigma_1^2>0,\sigma_2^2>0,\ |\rho|<1$.

(6) 均匀分布 $R(a,b)$ 或 $U(a,b),a<b$: 密度函数为

$$f(x;a,b)=(b-a)^{-1}I(a<x<b).$$

(7) 指数分布 $E(\alpha,\beta),\ -\infty<\alpha<\infty,\beta>0$: 密度函数为

$$f(x;\alpha,\beta)=\beta^{-1}\exp\{-(x-\alpha)/\beta\}I(x>\alpha),$$

其中 α 为位置参数, β 为刻度参数.

(8) 双指数分布(Laplace 分布) $DE(\alpha,\beta):\ -\infty<\alpha<\infty,\beta>0$: 密度函数为

$$f(x;\alpha,\beta)=(2\beta)^{-1}\exp\{-|x-\alpha|/\beta\},\quad -\infty<x<\infty.$$

(9) Cauchy 分布 $C(\alpha,\beta),-\infty<\alpha<\infty,\beta>0$: 密度函数为

$$f(x;\alpha,\beta)=\frac{\beta}{\pi[\beta^2+(x-\alpha)^2]},\quad -\infty<x<\infty.$$

(10) Logistic 分布 $L(\alpha,\beta),-\infty<\alpha<\infty,\ \beta>0$: 密度函数为

$$f(x;\alpha,\beta)=\frac{1}{\beta}\frac{\exp\{-(x-\alpha)/\beta\}}{\{1+\exp[-(x-\alpha)/\beta]\}^2},\quad -\infty<x<\infty.$$

(11) Gamma 分布 $G(\kappa,\beta),\kappa>0,\beta>0$, 密度函数为

$$f(x;\kappa,\beta)=\{\Gamma(\kappa)\beta^\kappa\}^{-1}x^{\kappa-1}\exp(-x/\beta)I(x>0),$$

其中 $\Gamma(\alpha)=\int_0^\infty t^{\alpha-1}\mathrm{e}^{-t}dt$. 当 $\alpha>0$ 时有 $\Gamma(\alpha+1)=\alpha\Gamma(\alpha)$, 且 $\Gamma(1/2)=\sqrt{\pi}$.

(12) Beta 分布 $Be(a,b),a>0,b>0$, 密度函数为

$$f(x;a,b)=\frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)}x^{a-1}(1-x)^{b-1}I(0<x<1).$$

设 X,Y 独立, $X\sim G(a,c),Y\sim G(b,c)$, 则 $V=X/(X+Y)\sim Be(a,b)$.

(13) 设 $X_0,X_1,\cdots,X_r$ 相互独立, $X_i\sim G(a_i,1),a_i>0,i=0,1,\cdots,r$. 记 $V_i=X_i/(X_0+X_1+\cdots+X_r)$. 此时, $(V_1,\cdots,V_r)$ 有密度函数

$$\begin{aligned}&f(v_1,\cdots,v_r;a_0,a_1,\cdots,a_r)\\=&\frac{\Gamma(a_0+\cdots+a_r)}{\Gamma(a_0)\cdots\Gamma(a_r)}\\&\times v_1^{a_1-1}\cdots v_r^{a_r-1}(1-v_1-\cdots-v_r)^{a_0-1}I(v_1>0,\cdots,v_r>0,\sum_{i=1}^{r}v_i<1).\end{aligned}$$

称 $(V_0,V_1,\cdots,V_r)$ 服从 Dirichlet 分布 $D(r;a_0,a_1,\cdots,a_r),a_0>0,\cdots,a_r>0$. Dirichlet 分布是 Beta 分布到多元的推广.

(14) Pareto 分布 $Pa(\zeta,\alpha),\zeta>0,\alpha>0$, 密度函数为

$$f(x;\zeta,\alpha)=\alpha\zeta^{\alpha}x^{-(\alpha+1)}I(x>\zeta).$$

易见, 若 $X\sim Pa(\zeta,\alpha)$, 则 $Y:=\log X\sim E(\log\zeta,1/\alpha)$.

其他一些重要分布, 还将在后面加以介绍.

1.3 统计推断与统计决策理论的基本概念

1.3.1 统计推断

1.2 节介绍了数理统计学的任务, 简言之, 包括收集数据和使用数据两个方面. 所谓"使用", 就是把按某种方式获取的样本, 用于解决所提统计问题. 正如 1.2 节所说, 一个确定的统计模型就是对样本分布族的一种界定, 它通常包含我们感兴趣的一些未知因素. 我们的任务就是要从样本中提取它所包含的分布的有用信息, 对分布族中的未知成分(例如未知参数)作出推断, 得出一定的结论, 这种结论或得出结论的过程都可称为**统计推断**. 统计推断就是要尽可能充分地利用这些信息, 对总体分布的未知成分作出尽可能精确可靠的论断.

"推断"一词, 一般是指从一定的条件和假定出发, 按照一定的方法或规则, 对未知事物得出的某种结论或得出结论的整个过程. 统计推断也是这种模式. 例如, 在参数分布族的情形, "未知事物" 就是未知参数, "一定的条件和假定" 就是样本和统计模型.

人们常把统计推断说成是由样本推断总体, 或由部分推断整体. 这是在掌握的知识不完全的情况下, 所做的一种归纳性的推理, 虽然不能保证在每一具体情况下推断的结果不会出错, 但使用概率论的方法, 可以给出种种有意义的指标, 对这种不确定性的程度给出确切的度量. 统计推断从结论的表述到优良性的刻画, 都离不开概

率, 这种表述和刻画, 也都具有确切的意义. 这种推断的意义正如在欧氏几何中推演出的任何一个命题那样明确. 随着以后章节的展开, 对此我们将会有更深的体会.

随着具体问题和统计模型的不同, 统计推断的形式也是多种多样的. 以正态总体 $N(\mu,\sigma^2)$ 中的问题为例, 设要估计均值 μ, 这时推断的形式是从样本 X 算出一个值, 以此作为 μ 的一个估计, 这种形式的推断叫做点估计. 我们也可以从样本 X 算出一个区间 $[\alpha(X),\beta(X)]$, 而推断 μ 在此区间内. 这种形式的推断叫做区间估计. 如果要求判断的是 "$\mu\leqslant 0$" 这一 " 假设" 是否成立, 则推断的形式是制定一个规则(检验), 依据这个规则, 对每个具体的样本 x 都能回答是应该接受还是否定这一假设. 这就是假设检验的问题. 以上是目前研究最深入的几种统计推断形式. 随着学习和实践的不断深入, 我们还会遇到其他一些重要的统计推断问题.

1.3.2 统计决策问题的三个要素

统计决策理论是 Wald 在 1950 年的著作 "Statistical Decision Function" 中创立的一种统计理论. 这个理论的基本观点已经渗入到不少统计分支中, 成为数理统计基础结构中的一个组成部分, 在应用上也有其重要性. 在此我们将介绍这一理论的一些基本概念.

按照 Wald 的理论, 要完整地描述一个统计问题, 不仅要给出统计模型, 还要明确需要解决的问题、解答采用的形式, 以及判定不同解法优劣的准则. 统计决策理论提供了解释这些问题的一种方便的框架. 在此框架下, 一个统计决策问题的描述包含以下三个要素.

1. 样本空间和样本分布族

这在 1.2 节已有介绍.

2. 决策空间 (或**行动空间**)

首先需要对解答采用何种形式加以界定. 对于所提的问题, 每一个具体的回答都称为一个 "决策" 或 "判决". 这种回答因问题的性质而异, 对点估计而言它是一个数, 对区间估计它是一个区间, 而对假设检验则是一个决定: 接受还是拒绝. 又因为一定的决策常导致采取一定的行动, 故也常把关于统计决策问题的回答称为一个 "行动". 对给定的问题而言, 所能作出的决策的全体是事先就明确的. 这种由所有可能的决策或行动组成的集合 $\mathcal{A}$ 就称为**决策空间**, 或**行动空间**.

除了上面提到的几种情况之外, 在应用上常见的一类问题是所谓 "多决策问题", 其决策空间 $\mathcal{A}$ 为一有限集, 元素个数一般大于 2. 如果要从 r 个处理中挑选一个最好的, 就会遇到这样的问题, 此时 $\mathcal{A}=\{a_1,\cdots,a_r\}$, 其中 a_i 表示 "认为第 i 个处理最好" 这一决定. 如果问题是要给这 r 个处理按优劣排个序, 则决策空间 $\mathcal{A}$ 包含 $r!$ 个元素, 每个元素都是 $(1,\cdots,r)$ 的一个置换.

为了理论上的需要, 我们还要引进可测空间 $(\mathcal{A},\mathcal{B}_{\mathcal{A}})$, 其中 $\mathcal{B}_{\mathcal{A}}$ 为 $\mathcal{A}$ 的某些子集构成的一个 σ-域.

3. 损失函数

在每一个具体问题中, 对于给定的分布参数 θ, 采用不同的决策或行动 a 会产生不同的后果, 我们必须对此进行比较. 为此, 可以假定这样的后果归结为损失 $L(\theta,a)$, L 为定义于 $\Theta\times\mathcal{A}$ 上的一个非负函数, 称为**损失函数**. 此外, 损失函数还需满足一个可测性条件, 即对每个给定的 θ, $L(\theta,\cdot)$ 为 $(\mathcal{A},\mathcal{B}_{\mathcal{A}})$ 上的一个可测函数.

在点估计的问题中, 假定 g 是未知的分布参数 θ 的一个已知函数, 要估计 $g(\theta)$. 假定决策空间为 $\mathcal{A}=(-\infty,\infty)$, 则损失函数可取为

$$L(\theta,a)=(g(\theta)-a)^2,$$

这就是在点估计理论中经常采用的平方损失函数. 而在一个形如

$$H_0:\theta\in\Theta_0 \quad \text{vs.} \quad H_1:\theta\in\Theta_1:=\Theta-\Theta_0$$

的假设检验问题中, 设决策空间 $\mathcal{A}=\{a_0,a_1\}$, 其中 a_0 和 a_1 分别表示接受和否定原假设, 则一个可供选择的损失函数为

$$L(\theta,a)=\begin{cases}0, & \text{当 } \theta\in\Theta_0,\ a=a_0,\ \text{或 } \theta\in\Theta_1,\ a=a_1,\\ b, & \text{当 } \theta\in\Theta_0,\ a=a_1,\\ c, & \text{当 } \theta\in\Theta_1,\ a=a_0.\end{cases}$$

此处, b,c 为两个适当选择的正数, 可用以区别两类错误的不同后果.

样本空间与样本分布族、决策空间、损失函数, 构成了一个统计决策问题的三个要素. 给出了这三个要素, 就给出了统计决策问题的完整提法.

1.3.3 统计决策函数及其风险函数

定义 1.5 设样本空间为 $(\mathcal{X},\mathcal{B})$, 决策空间为 $(\mathcal{A},\mathcal{B}_{\mathcal{A}})$. 定义于 $\mathcal{X}$ 上取值于 $\mathcal{A}$ 内的任一可测变换 δ, 都称为一个**统计决策函数**或**判决函数**.

对一个统计决策问题而言, 一旦根据某种考虑选定了一个决策函数 δ, 则每当有了样本 x, 即以 $\delta(x)$ 作为我们采取的决策或行动. 因此在 Wald 的理论中, 解决一个统计问题, 就是要选取一个适当的决策函数.

在一个具体问题中, 可供选择的决策函数很多. 如何比较这些决策函数的优劣呢? 设参数真值为 θ, 选用的决策函数为 δ, 则当样本为 X 时遭受的损失即为 $L(\theta,\delta(X))$. 在统计决策理论中, 用作比较的一个综合性指标, 即为当参数真值为 θ 时, 采用决策函数 δ 的平均损失, 它就是下面定义的风险函数.

定义 1.6 沿用上述记号, 且设 X 服从分布 P_θ, $\theta\in\Theta$, 则由下式定义的函数 $R(\cdot,\delta)$ 称为决策函数 δ 的**风险函数**,

$$R(\theta,\delta)=\mathrm{E}_\theta[L(\theta,\delta(X))]=\int_{\mathcal{X}}L(\theta,\delta(x))dP_\theta(x),\quad \theta\in\Theta. \tag{1.24}$$

上述定义的合理性, 是由于对给定的 θ, 由 $x \mapsto \delta(x)$ 和 $a \mapsto L(\theta, a)$ 的可测性, 立得复合函数 $x \mapsto L(\theta, \delta(x))$ 的可测性. 显然, 风险愈小, 决策函数就愈好. 为了对决策程序加以比较, 人们提出了若干优良性准则. 这些准则都是基于风险函数定义的, 将在稍后加以讨论.

随机化决策函数 在决策函数 δ 的上述定义中, 一旦有了样本 x, 所要采取的行动 $a = \delta(x)$ 就唯一地确定了. 但在一些情况下, 这个定义还需要加以拓展. 例如, 假定工厂向商店每月提供某种产品若干批, 约定每批产品都得经过抽样检验. 双方同意每批抽取10个产品, 若其中不合格品的数目 $x = 0$, 则商店应接受该批产品; 若 $x \geqslant 2$, 则应拒收. 而对 $x = 1$ 的情况, 双方意见不一致. 经协商, 双方同意当 $x = 1$ 时由投掷一枚硬币来决定, 若出现正面, 就接受该批产品, 否则就拒收. 即当某些样本 x 出现时, 所应采取的行动还不能完全定下来, 还要通过一个与 x 有关的随机机制去确定, 这样的决策函数称为随机化的. 将这种想法加以推广, 就可以给出随机化决策函数的定义.

定义 1.7 设样本空间为 $(\mathcal{X}, \mathcal{B})$, 决策空间为 $(\mathcal{A}, \mathcal{B}_{\mathcal{A}})$. 一个定义于 $\mathcal{X} \times \mathcal{B}_{\mathcal{A}}$ 上的非负函数 δ 称为**随机化决策函数**, 如果

(1) 对任何固定的 $A \in \mathcal{B}_{\mathcal{A}}$, $\delta(\cdot, A)$ 是一个 $\mathcal{B}$-可测函数;

(2) 对任何固定的 $x \in \mathcal{X}$, $\delta(x, \cdot)$ 是 $\mathcal{B}_{\mathcal{A}}$ 上的一个概率测度.

每当有了样本 x, 就依据概率分布 $\delta(x, \cdot)$ 从 $\mathcal{A}$ 中挑出一个元素 a, 作为采取的行动.

如果对每个 $x \in \mathcal{X}$, 分布 $\delta(x, \cdot)$ 都退化到某一点 a_x, 则回到原先的情形, 称之为**非随机化决策函数**.

下面推导**随机化决策函数的风险函数**. 为此, 首先算出在给定样本 x 的条件下的平均损失 $\int_{\mathcal{A}} L(\theta, a)\delta(x, da)$, 再对 x 求平均, 即得

$$R(\theta, \delta) = \mathrm{E}_\theta \left[\int_{\mathcal{A}} L(\theta, a)\delta(X, da) \right] = \int_{\mathcal{X}} \left\{ \int_{\mathcal{A}} L(\theta, a)\delta(x, da) \right\} dP_\theta(x), \ \theta \in \Theta. \quad (1.25)$$

由 $L(\theta, \cdot)$ 的可测性, 可知花括号中的积分有意义, 今证它作为 x 的函数, 是 $\mathcal{B}$-可测的, 因而外层积分也有意义. 为此, 只需证明, 对任一 $\mathcal{B}_{\mathcal{A}}$-可测的非负函数 f, 积分 $\int_{\mathcal{A}} f(a)\delta(x, da)$ 是 $\mathcal{B}$-可测的. 由随机化决策函数的定义, 易知当 f 为 $\mathcal{B}_{\mathcal{A}}$ 中任一集合的示性函数时, 上述结论成立. 用测度论中的标准方法, 知结论对非负简单函数, 进而对所有非负可测函数都成立.

随机化决策函数的引进, 其作用主要在理论方面, 即在求解某些统计问题时, 为了能达到某种优良性, 须采用这种决策函数. 一个熟知的例子是, 在对离散型分布的参数作假设检验时, 为精确达到某一指定的检验水平, 可能要采用随机化.

今设 δ_0 和 δ 为两个决策函数, 如果

$$R(\theta, \delta_0) \leqslant R(\theta, \delta), \quad \text{对} \ \forall\, \theta \in \Theta, \quad (1.26)$$

则称 δ_0 一致优于 δ. 此时如果至少存在一个 θ 使上式成立严格的不等号, 则称 δ_0 严格地一致优于 δ. 这时, 如果单从风险大小的角度考虑, 就没有理由舍弃 δ_0 而采用 δ. 这样, 在基于观测值作出统计决策时, 原先通过损失函数 $L(\theta,a)$ 选取行动 a 的问题, 就转化为通过风险函数 $R(\theta,\delta)$ 选取决策函数 δ 的问题了. 当然, 在实际使用中, 还常常需要考虑其他种种因素.

对于一个统计决策问题, 是否存在这样一个决策函数 δ_0, 使得 (1.26) 对任一决策函数 δ 都成立? 如是, 则称 δ_0 为一致最优的. 可惜的是, 在任何有意义的统计决策问题中, 这样的 δ_0 不可能存在. 考虑决策函数 $\delta(x)\equiv a_0$, 其中 a_0 为某个特定的决策. 假定对某个 θ_0, a_0 是正确的决策, 因而 $L(\theta_0,a_0)=0$, 则在 θ_0, δ 使风险取得最小值 0. 但可以想象, 采用这样的决策函数, 对于其他的 θ 值, 将会付出冒较大风险的代价. 由此可见, 通常不存在最优的决策程序, 可能是由于有些程序过多地注意了个别的参数值, 而忽视了整体的性能. 这就建议我们只限于考虑具有某种无偏性或对称性的决策程序, 可望在这样的范围内存在风险一致最小的程序. 这种类型的限制性条件, 例如无偏性、同变性、不变性, 将会在后面针对不同类型的统计问题提出来.

除了对程序的类加以限制之外, 另一种途径就是舍弃过于严格的一致最优性准则, 而建立其他一些较为宽松的整体性的优良性准则, 以期在这些准则之下可以找到最优解. 其中最常用的就是 Bayes 准则和 minimax 准则. Bayes 学派是现代统计中很有影响的一个学派, 有关这个学派的基本思想和方法, 我们将在后面予以讨论. 在此仅从优良性的角度作一点粗浅的说明. 简单地说, 就是取风险函数的某种加权平均作为这样一种准则. 为此, 可在可测空间 $(\Theta,\mathcal{B}_\Theta)$ 上取定一个概率分布 ξ, 通常称为**先验分布**. 对任一决策函数 δ, 定义其风险函数 $R(\theta,\delta)$ 对 $d\xi(\theta)$ 的加权平均 $R_\xi(\delta)=\int_\Theta R(\theta,\delta)d\xi(\theta)$, 作为评定决策程序优劣的一种准则, 这就是 **Bayes 准则**. 在上述准则下达到最优的决策函数, 称为在 "先验分布" ξ 下所给决策问题的 Bayes 解. 由于我们面对的是一个单指标优化问题, Bayes 解通常都存在. 至于 minimax 准则, 简言之, 则是从 "防止最坏的情况出现" 这一角度考虑, 取 $\sup_{\theta\in\Theta}R(\theta,\delta)$ 作为衡量决策程序 δ 的优良性的指标. 在考察的决策程序的类中, 使得这一指标达到最小的程序就称为 **minimax 程序**.

下面介绍与风险函数有关的两个重要概念.

容许性 对于一个决策函数 δ, 如果存在一个决策函数 δ^*, 其风险函数严格地一致优于 δ, 则称 δ 为非容许的, 否则就是容许的.

在数理统计学中, 通常有两类容许性问题, 一是证明特定的决策函数为容许 (或非容许), 二是寻求一特定问题中全部容许的决策函数. 容许性的研究是统计理论中最困难的课题之一, 至今进展有限且主要限于几类简单的参数估计模型. 我们在第二章中将对此作初步的介绍.

完全类和本质完全类 设 J 为一类决策函数, $J_1\subset J$. 若对任一 $\delta\in J-J_1$, 存在 $\delta_1\in J_1$ 严格地一致优于 δ, 则称 J_1 为 J 的一个完全子类. 如果 J 即为全体决策

函数的类, 则 J_1 简称为完全类. 若在上述定义中将 "严格地" 一词删去, 则 J_1 称为本质完全子类. Wald 决策理论的一个中心问题, 就是在一定的条件下, 确定某种特定的决策函数类是否为完全类或本质完全类.

在本节的末尾, 我们简单提一提统计推断和统计决策之间有些什么差别和联系. 一种看法认为, 它们的差别主要在于是否涉及后果, 认为统计推断是一种以探求未知为目的的单纯研究性的工作, 研究者只要考虑如何使其结论尽可能与实际符合, 不用考虑用在何处和效益如何. 而在统计决策问题中则引入了损失函数, 用以度量效益大小, 评价决策程序的优劣, 因此损失和风险是决策问题的出发点, 不同场合可能会有不同的考虑. 由此还产生一种看法, 认为在推断中不应掺入主观成分, 而决策问题的解, 因为与当事者的条件及掌握的信息有关, 故恰当的解不可能也不应排除主观因素. 至于统计决策问题与推断问题的联系, 从表面上看, 在于任何推断都可以赋予行动的意义. 实质上更重要的, 正如我们在后面的章节中会看到的, 统计决策的方法可以用来作为得出优良推断的一种手段. 换句话说, 尽管在推断问题中不涉及当事者的损失, 但这并不排斥把引进损失作为产生优良的统计推断方法的一种手段. 这样一来, 由于损失函数的多样性, 就可以提出大量优良性问题, 从而丰富了统计推断的理论和方法. 更为详细的评论, 可参阅陈希孺, 倪国熙 (1988).

1.4 统计量

1.4.1 定义和例子

1. 统计量

设 $(\mathcal{X},\mathcal{B})$ 为随机元 X 的样本空间, $(\mathcal{T},\mathcal{G})$ 为一可测空间. 如上所述, X 也可视为一个**样本**. 定义在 $\mathcal{X}$ 上而取值于 $\mathcal{T}$ 的可测函数 T 称为一个**统计量**. 统计量作为样本的函数, 可看作是对样本的一种加工. 在实际问题中, 作为随机试验或统计调查的原始资料而得到的样本, 常常是杂乱无章的, 用统计量可以将其加工成少数几个有代表性的数字, 它们集中反映了样本中所包含的我们感兴趣的信息. 因此, 统计量只依赖于样本而不依赖于任何未知的量, 尤其不能依赖于分布的未知参数.

在有些情况下, 我们只知道 T 定义于样本空间 $(\mathcal{X},\mathcal{B})$ 上而取值于空间 $\mathcal{T}$, 在空间 $\mathcal{T}$ 上并没有定义 σ-域. 令 $\mathcal{G}:=\{G\subset\mathcal{T}:T^{-1}(G)\in\mathcal{B}\}$, 则 $\mathcal{G}$ 为一 σ-域, 此时 T 即为 $(\mathcal{X},\mathcal{B})$ 到 $(\mathcal{T},\mathcal{G})$ 的一个可测变换, 也就是一个统计量. 易见, 刚刚定义的这个 σ-域是使 T 为可测而在 $\mathcal{T}$ 上定义的 σ-域中最大的一个.

有些统计量是由估计总体分布的某些数字特征产生的, 如各种样本矩、样本相关系数等. 有些统计量不是源于这一途径. 下面是一些例子.

例 1.3 设 $(\mathcal{X},\mathcal{B})$ 为样本空间, $X\sim dP_\theta(x)=p(x,\theta)d\mu(x)$, $\theta\in\Theta$, μ 为 $\mathcal{B}$ 上的一个 σ-有限测度. 每个样本 x 确定了 Θ 上的一个函数 $l_x(\cdot)=p(x,\cdot)$, 称之为似然函

数. 定义 $T(x)=l_x$. 为使 T 是一个统计量, 只要将 Θ 上所有非负实值函数的全体取为 $\mathcal{T}$, 它的 σ-域则取为 $\mathcal{G}=\{G\subset\mathcal{T}:T^{-1}(G)\in\mathcal{B}\}$, 此处 $T^{-1}(G)=\{x\in\mathcal{X}:l_x\in G\}$.

例 1.4 设 $X_1,\cdots,X_n$ 为取自某个一维总体的 i.i.d. 样本, 将其由小到大排列为 $X_{(1)}\leqslant X_{(2)}\leqslant\cdots\leqslant X_{(n)}$, 则 $(X_{(1)},\cdots,X_{(n)})$ 称为次序统计量. 设 $(X_1,\cdots,X_n)=(X_{(R_1)},\cdots,X_{(R_n)})$, 则称 $(R_1,\cdots,R_n)$ 为 $X_1,\cdots,X_n$ 的秩统计量, R_i 为 X_i 的秩. 若总体分布函数连续, 则当 $i\neq j$ 时 $P(X_i=X_j)=0$, 因而以概率 1 可将样本唯一地排成 $X_{(1)}<X_{(2)}<\cdots<X_{(n)}$ 的次序, 即次序统计量和秩统计量都是以概率 1 唯一确定的.

设总体指标 X 的分布函数为 F, $0<p<1$, 则称满足 $F(\nu-)\leqslant p\leqslant F(\nu)$ 的实数 ν 为总体 (或分布 F) 的一个 p-分位数, 记为 ξ_p. F 的 p-分位数一般不唯一, 其全体充满一个闭区间. $\xi_{1/2}$ 称为 X 或 F 的中位数. 不难证明, ν 为 X 的一个中位数, 当且仅当 $\mathrm{E}(|X-t|-|X|)$ 在 $t=\nu$ 处取最小值. 同样, ν 为 X 的一个 p-分位数, 当且仅当 $\mathrm{E}[\rho_p(X-t)-\rho_p(X)]$ 在 $t=\nu$ 处取最小值, 其中

$$\rho_p(z)=z[p-I(z<0)] \tag{1.27}$$

为一凸函数.

当总体 p-分位数唯一时, 可用样本 p-分位数去估计它. 样本 p-分位数可定义为极值问题

$$\sum_{i=1}^{n}\rho_p(X_i-t)=\min \tag{1.28}$$

的任何一个可测解, 它是 ξ_p 的一个 M-估计. 但在习惯上, 通常取 $X_{([(n+1)p])}$ 或 $X_{([np])}$ 作为样本 p-分位数, 此处 $[x]$ 表示不超过 x 的最大整数. 而样本中位数则取为 $X_{((n+1)/2)}$ 或 $\{X_{(n/2)}+X_{(n/2\,+1)}\}/2$, 视 n 为奇数或偶数而定.

例 1.5 设 $X_1,\cdots,X_n$ 为从某个一维总体中抽取的 i.i.d. 样本, 总体分布函数 F 连续. 定义

$$\mathbb{F}_n(x)=\sum_{i=1}^{n}I(X_i\leqslant x)/n,\qquad -\infty<x<\infty. \tag{1.29}$$

$\mathbb{F}_n$ 称为样本 $X_1,\cdots,X_n$ 的**经验分布函数**. 作为一个概率测度, $\mathbb{F}_n$ 在每个 X_i $(i\leqslant n)$ 处集中概率质量 $1/n$. 由 Glivenko-Cantelli 定理,

$$\sup_x|\mathbb{F}_n(x)-F(x)|\to 0\ \text{ a.s., 当 }\ n\to\infty.$$

对于给定的连续分布函数 F, 统计量 $\sqrt{n}\sup_x|\mathbb{F}_n(x)-F(x)|$ 的分布与 F 无关, 可用于检验数据跟 F 拟合的程度, 这就是著名的 Kolmogorov 统计量, 它的极限分布是 Kolmogorov 在 1933 年建立的.

2. 抽样分布

统计量 T 的分布称为 T 的**抽样分布**. 它取决于 X 的分布 P_θ 以及变换 T 的形

式, 可记为 P_θ^T:

$$P_\theta^T(G) = P_\theta\big[T^{-1}(G)\big], \quad \text{对 } \forall \ G \in \mathcal{G}. \tag{1.30}$$

但抽样分布的显式表达一般很难求得, 只有在少数情况下可以得到.

关于统计量和样本空间中的子 σ-域的关系, 以及积分变换的公式, 我们在 1.1.1 节中已经讨论过, 可参看引理 1.1 和引理 1.2. 关于指数族, 由引理 1.14 (i), 可以直接得到下述结果.

引理 1.15 设 X 的分布为指数族

$$dP_\theta(x) = C(\theta)\exp\left\{\sum_{i=1}^k \theta_i T_i(x)\right\} d\mu(x), \quad \theta \in \Theta, \tag{1.31}$$

$X_1, \cdots, X_n$ 为 X 的 i.i.d. 样本, 则 $T = \big(\sum_{j=1}^n T_1(X_j), \cdots, \sum_{j=1}^n T_k(X_j)\big)^\tau$ 的分布也是指数族, T 的分布可写成 $T \sim C^n(\theta)\exp(\theta^\tau t)d\nu(t)$ 的形式.

1.4.2 与正态样本有关的抽样分布

χ^2-分布、t-分布和 F-分布是在统计中有着广泛应用的三个分布, 有关内容在初等统计教科书中已有过介绍. 为了便于引用, 此处就有关的定义、分布的形式和有用的性质给出简单介绍.

χ^2-分布 设 $X_1, \cdots, X_n$ 相互独立, $X_i \sim N(a_i, 1)$, $i = 1, \cdots, n$. 令 $\xi = \sum_{i=1}^n X_i^2$, 则 ξ 的分布叫做自由度为 n、非中心参数为 $\delta := \big(\sum_{i=1}^n a_i^2\big)^{1/2}$ 的非中心 χ^2-分布, 记为 $\xi \sim \chi^2_{n,\delta}$. 当 $\delta = 0$ 时称之为中心的, 并简记为 χ^2_n.

t-分布 设 X, Y 独立, $X \sim \chi^2_n$, $Y \sim N(\delta, 1)$. 令 $\xi = Y/\sqrt{X/n}$, 则 ξ 的分布叫做自由度为 n、非中心参数为 δ 的非中心 t-分布, 记为 $\xi \sim t_{n,\delta}$. 注意此处 δ 可正可负. 当 $\delta = 0$ 时称之为中心的, 并简记为 t_n.

F-分布 设 X, Y 独立, $X \sim \chi^2_n$, $Y \sim \chi^2_{m,\delta}$. 令 $\xi = m^{-1}Y/(n^{-1}X)$, 则 ξ 的分布叫做自由度为 (m, n)、非中心参数为 δ 的 F-分布, 记为 $\xi \sim F_{m,n,\delta}$. 当 $\delta = 0$ 时称之为中心的, 并简记为 $F_{m,n}$.

为简便起见, 在提及给定自由度的中心的 χ^2-变量、t-变量和 F-变量时, 可省去 "中心的" 一词.

为求这些分布的密度函数, 关键在于求 χ^2-分布的密度函数. 一旦求出了 χ^2-分布的密度函数, t-分布和 F-分布的密度函数就可通过商的密度的常规计算得到. 此处的密度都是相对于 Lebesgue 测度而言的!

此处我们不加证明地引述以下引理.

引理 1.16 设 $X_1, \cdots, X_n$ 相互独立, $X_i \sim N(a_i, 1)$, $i = 1, \cdots, n$. 作正交变换 $Y_i = \sum_{j=1}^n c_{ij}X_j$, $i = 1, \cdots, n$, 则 $Y_1, \cdots, Y_n$ 相互独立, $Y_i \sim N(b_i, 1)$, $i = 1, \cdots, n$, 即独立等方差的正态变量, 经正交变换后仍保有这些性质.

为了方便, 我们再次介绍 Gamma 函数和 Beta 函数如下:

$$\Gamma(a)=\int_0^\infty x^{a-1}\mathrm{e}^{-x}dx,\ a>0; \tag{1.32}$$

$$\beta(a,b)=\int_0^1 t^{a-1}(1-t)^{b-1}dt,\ a>0,\ b>0. \tag{1.33}$$

容易验证, 当 $a>0$ 时, $\Gamma(a+1)=a\Gamma(a)$, $\Gamma(1/2)=\sqrt{\pi}$, $2^{2i}i!\Gamma(i+1/2)=\sqrt{\pi}(2i)!$, 且 $\beta(a,b)=\Gamma(a)\Gamma(b)\big/\Gamma(a+b)$.

首先要证明, 非中心 χ^2-分布的密度函数的表达式只依赖于 n 和 δ. 为此, 作正交变换 $Y=CX$, 其中 $Y=(Y_1,\cdots,Y_n)^\tau$, $X=(X_1,\cdots,X_n)^\tau$, 而正交阵 C 的第一行为 $(a_1/\delta,\cdots,a_n/\delta)$. 按引理 1.16, $Y_1,\cdots,Y_n$ 相互独立, $Y_i\sim N(b_i,1)$, $i=1,\cdots,n$, 且若记 $a=(a_1,\cdots,a_n)^\tau$, $b=(b_1,\cdots,b_n)^\tau$, 有 $b=Ca$. 由 C 的取法及其行向量两两正交的性质, 有 $b_2=\cdots=b_n=0$, $b_1=\delta$. 又因正交变换保持距离不变, 故 $\xi=\sum_{i=1}^n X_i^2=\sum_{i=1}^n Y_i^2$, 因而 ξ 的分布只依赖于 n 和 δ.

记 $\zeta=\sum_{i=2}^n Y_i^2$, 则 $\xi=Y_1^2+\zeta$, 其中 $Y_1\sim N(\delta,1)$, $\zeta\sim\chi^2_{n-1}$ 且两者独立. 用 K_n 记 χ^2_n 的分布函数. 依定义, 当 $x\leqslant 0$ 时, $K_n(x)=0$; 而当 $x>0$ 时, 有

$$K_n(x)=(2\pi)^{-n/2}\int_{||t||\leqslant\sqrt{x}}\exp(-||t||^2/2)dt_1\cdots dt_n.$$

变换到球坐标, 得

$$K_n(x)=c_n\int_0^{\sqrt{x}} r^{n-1}\exp(-r^2/2)dr,$$

此处, 常数 c_n 可由 $K_n(\infty)=1$ 确定,

$$c_n=\{2^{n/2-1}\Gamma(n/2)\}^{-1}.$$

由此可确定 $K_n(x)$ 进而即得 χ^2_n 的密度函数

$$k_n(x)=\frac{1}{2^{n/2}\Gamma(n/2)}x^{n/2-1}\mathrm{e}^{-x/2}I(x>0). \tag{1.34}$$

另一方面, Y_1^2 有密度

$$\begin{aligned}g(x)&=\frac{1}{\sqrt{8\pi x}}\Big\{\exp\Big[-\frac{(\sqrt{x}-\delta)^2}{2}\Big]+\exp\Big[-\frac{(\sqrt{x}+\delta)^2}{2}\Big]\Big\}I(x>0)\\&=\frac{1}{\sqrt{2\pi}}\sum_{j=0}^\infty\frac{\delta^{2j}}{(2j)!}\exp(-\delta^2/2)x^{j-1/2}\exp(-x/2)I(x>0).\end{aligned}$$

由 (1.34), 上式可写成

$$g(x)=\sum_{j=0}^\infty\frac{1}{j!}(\delta^2/2)^j\exp(-\delta^2/2)k_{2j+1}(x).$$

设 $J\sim\mathrm{Poisson}(\delta^2/2)$, 且与 ζ 独立, 则上式表明当 $J=j$ 给定时, Y_1^2 的条件分布为 χ^2_{2j+1}, 而 ξ 的条件分布则为 χ^2_{2j+n}. 由此立得 ξ 即 $\chi^2_{n,\delta}$ 的密度函数

$$k_{n,\delta}(x)=\sum_{j=0}^\infty\frac{(\delta^2/2)^j\exp(-\delta^2/2)}{j!2^{j+n/2}\Gamma(j+n/2)}x^{j+n/2-1}\exp(-x/2)I(x>0). \tag{1.35}$$

由上述推导及 1.2.3 小节中介绍的有关分布, 我们便可得到如下推论.

推论 1.17 (1) $\chi_n^2 \overset{d.}{=} G(n/2,2)$.

(2) 若 $Z_1,\cdots,Z_n$ 为 i.i.d. $N(0,1)$ 变量, 则 $Z_1^2/(Z_1^2+\cdots+Z_n^2)\sim Be\big(1/2,(n-1)/2\big)$.

(3) 设 $e_1,\cdots,e_n$ i.i.d., $e_1\sim E(0,\beta)$, 则 $2\sum_{j=1}^n e_j/\beta \sim \chi_{2n}^2 \overset{d.}{=} G(n,2)$.

(4) 设 $J\sim \text{Poisson}(\delta^2/2)$, 且 $\xi\big|J=j\sim\chi_{2j+n}^2$, 则 $\xi\sim\chi_{n,\delta}^2$.

$\chi_{n,\delta}^2$ 的特征函数为 $\psi(t)=(1-2\mathrm{i}t)^{-n/2}\exp\{\mathrm{i}\delta^2 t/(1-2\mathrm{i}t)\}$, 此处 $\mathrm{i}=\sqrt{-1}$.

今设 $X>0$, X,Y 独立, 分别有密度 f_1 和 f_2, 则商 Y/X 的密度为 $g(x)=\int_0^\infty tf_1(t)f_2(tx)dt$. 由此及 (1.35), 知 $t_{n,\delta}$ 的密度为

$$h_{n,\delta}(x)=\frac{\exp(-\delta^2/2)}{\sqrt{n\pi}\Gamma(n/2)}\sum_{j=0}^\infty\Big(\frac{2}{n}\Big)^{j/2}\Gamma\Big(\frac{n+j+1}{2}\Big)\frac{(\delta x)^j}{j!}\Big(1+\frac{x^2}{n}\Big)^{-(n+j+1)/2}. \tag{1.36}$$

t_n 的密度为

$$h_n(x)=\frac{\Gamma\big((n+1)/2\big)}{\sqrt{n\pi}\Gamma(n/2)}\Big(1+\frac{x^2}{n}\Big)^{-(n+1)/2}. \tag{1.37}$$

$F_{m,n,\delta}$ 的密度为

$$f_{m,n,\delta}(x)=\mathrm{e}^{-\delta^2/2}\sum_{j=0}^\infty\frac{(\delta^2/2)^j}{j!}\frac{\Gamma\big((m+n)/2+j\big)}{\Gamma(n/2)\Gamma(m/2+j)}\frac{n^{n/2}m^{m/2+j}x^{m/2+j-1}}{(n+mx)^{(m+n)/2+j}}I(x>0). \tag{1.38}$$

而 $F_{m,n}$ 的密度为

$$f_{m,n}(x)=\frac{\Gamma\big((m+n)/2\big)}{\Gamma(n/2)\Gamma(m/2)}\frac{n^{n/2}m^{m/2}x^{m/2-1}}{(n+mx)^{(m+n)/2}}I(x>0). \tag{1.39}$$

在统计推断中常常用到这些分布的分位数. χ_n^2 的 $1-\alpha$ 分位数, 即 χ_n^2 的上 α 分位数, 记为 $\chi_n^2(\alpha)$. 对于其他分布, 也有类似的记号.

下面列出正态样本的一些重要统计量的抽样分布, 它们都属于上述三大分布. 证明都不难, 在初等统计的教科书中也可以找到.

(i) 设 $X_1,\cdots,X_n$ 为从正态总体 $N(a,\sigma^2)$ 中抽出的 i.i.d. 样本, $n\geqslant 2$. 记 $\bar X=\sum_{j=1}^n X_j/n$, $S^2=\sum_{j=1}^n(X_j-\bar X)^2/(n-1)$, 分别称之为样本均值和样本方差, 则 $\bar X$ 与 S^2 独立, 且 $\bar X\sim N(a,\sigma^2/n)$, $(n-1)S^2/\sigma^2\sim\chi_{n-1}^2$. 设 b 为一已知的实数, 则 $\xi:=\sqrt{n}(\bar X-b)/S\sim t_{n-1,\delta}$, 其中 $\delta=\sqrt{n}(a-b)/\sigma$ 为非中心参数. ξ 称为一样本 t-统计量.

为证此结论, 只需对 $X=(X_1,\cdots,X_n)^\tau$ 作正交变换 $Y=CX$, 而正交阵 C 的第一行的 n 个元素均取为 $n^{-1/2}$, 并利用引理 1.16.

(ii) 设 $X_1,\cdots,X_n$ 和 $Y_1,\cdots,Y_m$ 分别是从总体 $N(a_1,\sigma^2)$ 和 $N(a_2,\sigma^2)$ 中抽出的 i.i.d. 样本, 且两样本相互独立. 令

$$\bar X=\sum_{j=1}^n X_j/n,\bar Y=\sum_{j=1}^m Y_j/m,S^2=\Big\{\sum_{j=1}^n(X_j-\bar X)^2+\sum_{j=1}^m(Y_j-\bar Y)^2\Big\}/(m+n-2),$$

则

$$\xi := \sqrt{mn/(m+n)}(\bar{Y} - \bar{X} - b)/S \sim t_{m+n-2,\delta},$$

其中 $\delta = \sqrt{mn/(m+n)}(a_2 - a_1 - b)/\sigma$. 本例中的 ξ 称为两样本 t-统计量.

为了证明, 记 $Z = (X_1, \cdots, X_n, Y_1, \cdots, Y_m)^\tau, W = (W_1, \cdots, W_{m+n})^\tau$. 作正交变换 $W = CZ$, 其中正交阵 C 的第一、二行分别取为 $(n^{-1/2}, \cdots, n^{-1/2}, 0, \cdots, 0)$ 和 $(0, \cdots, 0, m^{-1/2}, \cdots, m^{-1/2})$, 则

$$(m+n-2)S^2 = \sum_{i=3}^{m+n} W_i^2,\ \bar{Y} - \bar{X} = W_2/\sqrt{m} - W_1/\sqrt{n}.$$

由引理 1.16, 不难得到上述结果.

(iii) 设样本同上, 则

$$\xi := \frac{1}{m-1}\sum_{j=1}^{m}(Y_j - \bar{Y})^2 \Big/ \left[\frac{1}{n-1}\sum_{j=1}^{n}(X_j - \bar{X})^2\right] \sim F_{m-1,n-1}.$$

ξ 称为 F-统计量.

1.4.3 对称幂等方阵与 χ^2 分布

正态向量的二次型在一些统计问题中有重要应用. 本小节讨论它们与 χ^2-分布的关系. 为此, 设 $X_1, \cdots, X_n$ 相互独立, $X_i \sim N(a_i, 1)$, $i = 1, \cdots, n$. 记 $X = (X_1, \cdots, X_n)^\tau, a = (a_1, \cdots, a_n)^\tau$. 此处, 我们不加证明地引述下列结果.

(i) 设 A 为 n 阶对称方阵, 则 $Y = X^\tau AX$ 服从 χ^2-分布的充要条件为 A 为幂等方阵, 即 $A^2 = A$. 此时, $Y \sim \chi^2_{r,\delta}$, 其中 $r = \mathrm{rank}(A)$, 而 $\delta^2 = a^\tau Aa$.

(ii) 设 A 和 A_1 都是 n 阶对称方阵, $Y = X^\tau AX \sim \chi^2_{m,\delta}, Y_1 = X^\tau A_1 X \sim \chi^2_{m_1,\delta_1}$, $A_2 := A - A_1$ 为非零的半正定方阵, 则 $m_2 := m - m_1 > 0$, $\delta_2^2 := \delta^2 - \delta_1^2 \geqslant 0$, $Y_2 := Y - Y_1 = X^\tau A_2 X \sim \chi^2_{m_2,\delta_2}$, 且 Y_1 与 Y_2 独立, $A_1 A_2 = 0$.

(iii) (Cochran 定理) 设 $A_1, \cdots, A_m$ 都是 n 阶对称方阵, $X^\tau X = \sum_{i=1}^m X^\tau A_i X$, 则 $X^\tau A_1 X, \cdots, X^\tau A_m X$ 均为 χ^2-变量且相互独立的充要条件为 $\sum_{i=1}^m \mathrm{rank}(A_i) = n$. 当该条件成立时, 有 $X^\tau A_i X \sim \chi^2_{n_i,\delta_i}, n_i = \mathrm{rank}(A_i), \delta_i^2 = a^\tau A_i a$.

有关证明可参看陈希孺 (1981, 1997) 和 (1999, 2009).

1.5 充分统计量

统计量的充分性是统计学中最基本的概念之一, 是 R. A. Fisher 引进 (1920) 和命名 (1922) 的. 在他 1922 年这一奠基性的论文中, Fisher 还陈述了因子分解准则. 在一般的控制族中这一准则的证明是由 Halmos 和 Savage (1949) 完成的. 有关内容将在本节加以讨论.

1.5.1 定义和例子

设 X 的样本空间为 $(\mathcal{X},\mathcal{B})$, 其分布族为 $\mathcal{P}=\{P_\theta,\theta\in\Theta\}$, Θ 是参数空间. 在将样本 X 加工成统计量 T 时, 数据中包含的有关未知分布的信息可能会有所损失. 我们自然希望这种损失越小越好. 有一种数据处理的方法特别值得重视, 它本质上在于仅仅舍弃了数据中不含未知分布信息的部分, 这些部分对关于 θ 的任何决策问题都是毫无价值的.

例 1.6 设在 n 次 (独立的) Bernoulli 试验中, 事件 S (成功) 每次发生的概率 p 未知. 令 $X_i=1$ 或 0, 视第 i 次试验结果是否为成功而定. 样本 $(X_1,\cdots,X_n)$ 显示了有多少次成功, 以及它们在哪些试验发生. 这第二部分信息, 无助于对 p 的了解. 一旦成功总次数 $\sum X_i$ 已知并等于 t, 则不管 p 取什么值, 这些成功发生在 $\binom{n}{t}$ 个可能位置的每一个都具有相同的条件概率. 因此, 尽管既不知道每个 X_i 又不知道 p, 只要知道了 $\sum X_i$, 就可借助于一个随机数字表构造出一组随机变量 $X_1^*,\cdots,X_n^*$, 使之与 $X_1,\cdots,X_n$ 有相同的联合分布. 因此, 包含在 $(X_1,\cdots,X_n)$ 中的信息, 跟包含在 $\sum X_i$ 和一个随机数字表中的信息完全相同.

例 1.7 设 $X_1,\cdots,X_n$ i.i.d., $X_1\sim N(0,\sigma^2)$. 在每个球面 $\sum X_i^2=\text{constant}$ 上, 样本点的条件分布是球面上的均匀分布, 且与 σ^2 无关. 由此可构造一个等价的样本 $X_1^*,\cdots,X_n^*$, 只要用到 $\sum X_i^2$ 的知识, 和可以在球面上随机取点的一种机械.

粗略地说, 称统计量 T 关于分布族 $\mathcal{P}$ (或关于 θ) 是充分的, 若对给定的 $T=t$, X 的条件分布与 θ 无关. 正如上面两个例子所显示的, 如果只观察了 T 而不是 X, 则通过 T 的观测和一个随机化机制, 即可定义一个随机变量 X^*, 它与 X 具有相同的分布. 这样, 给定基于 X 的任何决策程序, 总可以构造一个基于 X^* 的等价程序, 它可视为单纯依赖于 T 的一个随机化程序. 因此, 只要允许采用随机化程序, 即使只限于考虑充分统计量, 也不会失去一般性.

下面我们对充分统计量给出一个更为一般的严格定义.

定义 1.8 设 $\mathcal{P}=\{P_\theta,\theta\in\Theta\}$ 是定义在样本空间 $(\mathcal{X},\mathcal{B})$ 上的一族分布, 称统计量 T 关于 $\mathcal{P}$ (或关于 θ)是充分的, 如果对每个 $B\in\mathcal{B}$, 存在条件概率 $P_\theta(B|T)$ 的一个与 θ 无关的版本 $P(B|T)$.

上述定义也可由参数族推广到一般的分布族 $\mathcal{P}$. 值得注意的是, 此定义对条件概率分布的存在性并无要求.

设 $X_1,\cdots,X_n$ 是独立同分布 (i.i.d.) 的随机变量, 且具有连续的分布函数 $F_\theta,\theta\in\Theta$, 则由例 1.1, 次序统计量 $T=(X_{(1)},\cdots,X_{(n)})$ 关于 θ 是充分的.

例 1.8 设 $X_1,\cdots,X_n$ 是从指数分布 $E(0,b),b>0$ 的总体中取出的 i.i.d. 样本, 则 $T:=X_1+\cdots+X_n$ 为充分统计量. 为证此结论, 令 $T=X_1+\cdots+X_n,V_i=X_i/(X_1+\cdots+X_n)$, $i\leqslant n$. 将其视为 $X_1,\cdots,X_n$ 到 $(T,V_2,\cdots,V_n)$ 之间的变换, 这个变换是一对一的. 我们只需证明当 T 给定时, $(V_2,\cdots,V_n)$ 的条件密度与 b 无关.

经过计算, $(T, V_2, \cdots, V_n)$ 的联合密度为

$$g(t; v_2, \cdots, v_n) = \big(\Gamma(n)b^n\big)^{-1} t^{n-1}\mathrm{e}^{-t/b} I(t>0) \times \Gamma(n) I(v_2>0, \cdots, v_n>0, v_2+\cdots+v_n<1).$$

由此可知, $(V_1, \cdots, V_n)$ 与 T 独立, 其分布为与 b 无关的 Dirichlet 分布 $D(n; 1, \cdots, 1)$ (参看 1.2.3 小节).

在欧氏样本空间的情形, 我们有下述定理.

定理 1.18 设样本空间 $(\mathcal{X}, \mathcal{B})$ 是欧氏的. 如果统计量 $T: \mathcal{X} \mapsto \mathcal{T}$ 关于分布族 $\mathcal{P} = \{P_\theta\}$ 是充分的, 则存在条件概率分布的与 θ 无关的一个版本 $P_\theta(B|t) = Q(B, t)$, 对每个 $t \in \mathcal{T}$, $Q(B, t)$ 是 $\mathcal{B}$ 上的一个概率测度.

证明 这可从定理 1.7 的证明中看出. 由充分性的定义, 对每个有理数 r, 可取函数 $F(r, t)$ 与 θ 无关, 因而最后得到的条件分布也与 θ 无关. ■

1.5.2 因子分解定理

直接用定义去验证统计量的充分性是困难的. 所幸的是, 下面的因子分解定理提供了一个涵盖面很广且简便的充分性判别准则. 由于定理和有关引理的证明有一定难度, 这些证明在初读时可以略去.

定理的建立要借助于测度受控族的一个引理. 为此, 我们给出如下的定义. 称可测空间 $(\mathcal{X}, \mathcal{B})$ 上的一个测度族 $\mathcal{N}$ 受控于一个 σ-有限测度 μ, 如果对任何 $\nu \in \mathcal{N}$, 都有 $\nu \ll \mu$, 记为 $\mathcal{N} \ll \mu$. 如果这样的一个 μ 存在, 则称测度族 $\mathcal{N}$ 为一个测度受控族.

$(\mathcal{X}, \mathcal{B})$ 上的两个测度族 $\mathcal{M}$ 和 $\mathcal{N}$ 称为等价的, 如果对任何 $B \in \mathcal{B}$, $\mu(B) = 0$ 对所有$\mu \in \mathcal{M} \Longrightarrow \nu(B) = 0$ 对所有 $\nu \in \mathcal{N}$, 并且反之亦真.

引理 1.19 $(\mathcal{X}, \mathcal{B})$上的一族概率测度 $\mathcal{P}$ 被一个 σ-有限测度 μ 所控制, 当且仅当 $\mathcal{P}$ 有可数的等价子类.

证明 假定 $\mathcal{P}$ 有可数的等价子类 $\{P_1, P_2, \cdots\}$, 则 $\mathcal{P}$ 被 $\mu := \sum_n P_n/2^n$ 所控制.

反过来, 设 $\mathcal{P}$ 被 σ-有限测度 μ 所控制, 不失一般性, 可以假定 μ 为有限.

记 $\mathcal{Q} = \{Q = \sum_i c_i P_i : \sum c_i = 1, c_i > 0, P_i \in \mathcal{P}, i = 1, 2, \cdots\}$. 显然 $\mathcal{Q} \ll \mu$. 令 $q = dQ/d\mu$. 往证: $\exists\, Q_0 \in \mathcal{Q}$, 使对 $\forall B \in \mathcal{B}$, $Q_0(B) = 0 \Longrightarrow Q(B) = 0$ 对 $\forall\, Q \in \mathcal{Q}$.

令 $\mathcal{C} = \{C \in \mathcal{B} : \exists\, Q \in \mathcal{Q},$ 使 $Q(C) > 0,$且在 C 上, $q(x) > 0, \text{a. e. } \mu\}$. 显然 $\mathcal{C}$ 非空; 且若 $C_1, C_2 \in \mathcal{C}, C_1 \cap C_2 = \emptyset$, 则 $C_1 \cup C_2 \in \mathcal{C}$. 取 $C_i \in \mathcal{C}$, 使 $\mu(C_i) \to \sup_{\mathcal{C}} \mu(C)$. 不妨设 $q_i(x) = dQ_i/d\mu$ 在 C_i 上处处大于 0. 记 $C_0 = \bigcup_i C_i$, $Q_0 = \sum_{i=1}^{\infty} Q_i/2^i$, 则在 C_0 上 $q_0(x) > 0$, 故 $Q_0 \in \mathcal{C}$. 今证 Q_0 即为所求.

为此, 设 $Q_0(B) = 0$. 任取 $Q \in \mathcal{Q}$, 由 $Q_0(B \cap C_0) = 0$ 可得 $\mu(B \cap C_0) = 0$ 以及 $Q(B \cap C_0) = 0$. 令 $C = \{x : q(x) > 0\}$, 则 $Q(B \cap C_0^c \cap C^c) = 0$. 最后令 $E := B \cap C_0^c \cap C$, 如果 $Q(E) > 0$, 则 $E \in \mathcal{C}$, 因而 $C_0 \cup E \in \mathcal{C}$, 且 $\mu\{C_0 \cup E\} > \mu(C_0)$, 与 $\mu(C_0) = \sup_{\mathcal{C}} \mu(C)$ 矛盾. 由此即得 $Q(E) = 0$, 因而 $Q(B) = 0$.

设 $Q_0=\sum_i c_iP_i$, 则 $\{P_1,P_2,\cdots\}$ 即为 $\mathcal{P}$ 的一个可数的等价子类. ■

定理 1.20 (因子分解定理) 设 X 的样本空间为 $(\mathcal{X},\mathcal{B})$,

$$X\sim dP_\theta(x)=p_\theta(x)d\mu(x),\quad \theta\in\Theta,$$

其中 μ 为 $\mathcal{B}$ 上的一个 σ-有限测度. 又设 T 为取值于可测空间 $(\mathcal{T},\mathcal{G})$ 的一个统计量, 则 T 关于 $\mathcal{P}\equiv\{P_\theta,\theta\in\Theta\}$ 为充分的, 当且仅当对每个固定的 θ, 存在 $\mathcal{T}$ 上的一个非负 $\mathcal{G}$-可测函数 g_θ 和 $\mathcal{X}$ 上的一个非负 $\mathcal{B}$-可测函数 h, 使得

$$p_\theta(x)=g_\theta\big(T(x)\big)h(x)\quad \text{a.e. } \mu. \tag{1.40}$$

上式中的例外集可以与 θ 有关.

定理证明基于引理中已经证明的事实: 存在概率分布 $\lambda=\sum_i c_iP_{\theta_i}$ $(\forall\theta_i\in\Theta)$ 与分布族 $\mathcal{P}$ 等价, 即

$$\text{对任何 } B\in\mathcal{B},\lambda(B)=0 \text{ 当且仅当对任何 } \theta\in\Theta \text{ 有 } P_\theta(B)=0. \tag{1.41}$$

我们可以先建立以下的定理.

定理 1.21 设 $\mathcal{P}=\{P_\theta,\theta\in\Theta\}$ 为 $(\mathcal{X},\mathcal{B})$ 上概率分布的一个受控族, $\lambda=\sum c_iP_{\theta_i}$ 满足 (1.41), 则取值于可测空间 $(\mathcal{T},\mathcal{G})$ 的统计量 T 关于 $\mathcal{P}$ 是充分的, 当且仅当对每个固定的 θ, 存在一个非负 $\mathcal{G}$-可测函数 g_θ, 使得

$$dP_\theta(x)=g_\theta\big(T(x)\big)d\lambda(x). \tag{1.42}$$

证明 设 T 关于 θ 是充分的, 则对任一 $B\in\mathcal{B}$, 存在条件概率 $P_\theta[B|T(x)]$ 的一个与 θ 无关的版本 $P[B|T(x)]$. 由 (1.7), 对所有 $\theta\in\Theta,B_0\in\mathcal{B}_0:=T^{-1}(\mathcal{G})$,

$$\int_{B_0}P\big[B|T(x)\big]dP_\theta(x)=P_\theta(B\cap B_0), \tag{1.43}$$

由于 $\lambda:=\sum_i c_iP_{\theta_i}$ $(\forall\ \theta_i\in\Theta)$, 有

$$\int_{B_0}P\big[B|T(x)\big]d\lambda(x)=\lambda(B\cap B_0),$$

即 $P\big[B|T(x)\big]$ 可取作事件 B 在概率 λ 下的条件概率函数 $P_\lambda[B|T(x)]$.

由 Radon-Nikodym 定理, 存在一个 $\mathcal{G}$-可测函数 g_θ, 使 $g_\theta\big(T(x)\big)$ 为 P_θ 对 $(\mathcal{B}_0,\lambda)$ 的 Radon-Nikodym 导数. 对任一 $B\in\mathcal{B}$, 有

$$\begin{aligned}
P_\theta(B)&=\int P\big[B|T(x)\big]dP_\theta(x) && \text{由 (1.43)}\\
&=\int \mathrm{E}_\lambda\big[I_B(X)|T(x)\big]dP_\theta(x) && \text{由 } P_\lambda[B|\ T(x)]=P[B|T(x)]\\
&=\int \mathrm{E}_\lambda[I_B(X)|T(x)]g_\theta\big(T(x)\big)d\lambda(x) && \text{由 } dP_\theta(x)/d\lambda\big|_{\mathcal{B}_0}=g_\theta\big(T(x)\big)\\
&=\int \mathrm{E}_\lambda\big\{g_\theta\big(T(X)\big)I_B(X)|T(x)\big\}d\lambda(x) && \text{由引理 1.5 (iii)}\\
&=\int g_\theta\big(T(x)\big)I_B(x)d\lambda(x) && \text{由条件期望的平滑性}\\
&=\int_B g_\theta\big(T(x)\big)d\lambda(x),
\end{aligned}$$

即 (1.42) 成立.

今设 (1.42) 成立, 要证对任一给定的 θ 和 $B\in\mathcal{B}$, $P_\theta[B|T(x)]=P_\lambda[B|T(x)]$. 由 (1.42), 对任一 $B_0\in T^{-1}(\mathcal{G})=\mathcal{B}_0$,

$$P_\theta(B\cap B_0)=\int_{B_0}P_\theta[B|T(x)]dP_\theta(x)=\int_{B_0}P_\theta[B|T(x)]g_\theta\big(T(x)\big)d\lambda(x);$$

另一方面, 由 (1.42), 并考虑到 $\mathrm{E}_\lambda[\cdot\,|T(x)]$ 的定义, 有

$$\begin{aligned}P_\theta(B\cap B_0)&=\int_{B_0}I_B(x)dP_\theta(x)=\int_{B_0}I_B(x)g_\theta\big(T(x)\big)d\lambda(x)\\&=\int_{B_0}\mathrm{E}_\lambda\big[I_B(X)g_\theta\big(T(X)\big)|T(x)\big]d\lambda(x)\\&=\int_{B_0}P_\lambda\big[B|T(x)\big]g_\theta\big(T(x)\big)d\lambda(x).\end{aligned}$$

故对任一 $B_0\in\mathcal{B}_0$, $\int_{B_0}P_\theta[B|T(x)]g_\theta\big(T(x)\big)d\lambda(x)=\int_{B_0}P_\lambda\big[B|T(x)\big]g_\theta\big(T(x)\big)d\lambda(x)$, 故 $P_\theta\big[B|T(x)\big]g_\theta\big(T(x)\big)=P_\lambda\big[B|T(x)\big]g_\theta\big(T(X)\big)$ a.e. $(\mathcal{B}_0,\lambda)$ 因而 a.e. $(\mathcal{B}_0,P_\theta)$ 成立. 由 (1.42), $g_\theta\big(T(x)\big)\neq0$ a.e. $(\mathcal{B}_0,P_\theta)$, 由此得出 $P_\theta\big[B|T(x)\big]=P_\lambda\big[B|T(x)\big]$ a.e. $(\mathcal{B}_0,P_\theta)$, 这证明了 T 对 θ 的充分性. ■

现在回到因子分解定理的证明. 为此, 设 $\lambda=\sum_i c_iP_{\theta_i}$ 满足 (1.41). 如果 T 是充分的, 则 (1.40) 可由 (1.42) 得到, 其中 $h=d\lambda/d\mu$. 反过来, 如果 (1.40) 成立, 则

$$d\lambda(x)=\sum c_ig_{\theta_i}\big(T(x)\big)h(x)d\mu(x):=\psi\big(T(x)\big)h(x)d\mu(x).$$

因而 $dP_\theta(x)=g^*_\theta\big(T(x)\big)d\lambda(x)$, 此处, 当 $\psi(t)>0$时, $g^*_\theta(t)=g_\theta(t)/\psi(t)$, 而当 $\psi(t)=0$ 时 $g^*_\theta(t)$ 可取任意值. 定理证毕.

例 1.9 设 $X_1,\cdots,X_n$ 为取自指数族 (1.12) 的 i.i.d. 样本, 则其联合分布可表示为

$$dP_\theta(x_1,\cdots,x_n)=C^n(\theta)\exp\Big\{\sum_{i=1}^kQ_i(\theta)\sum_{j=1}^nT_i(x_j)\Big\}\prod_{j=1}^n\big\{h(x_j)d\mu(x_j)\big\}.$$

由因子分解定理, $T:=\big(\sum_{j=1}^nT_1(X_j),\cdots,\sum_{j=1}^nT_k(X_j)\big)$ 为充分统计量.

上述论断适用于正态、二项、Poisson 等常见分布族. 例如, 设 $X_1,\cdots,X_n$ i.i.d., $X_1\sim N(a,\sigma^2),-\infty<a<\infty,\ \sigma^2>0$. 记

$$\bar X=\sum_{i=1}^nX_i/n,\ S^2=\sum_{i=1}^n(X_i-\bar X)^2/(n-1),$$

则由 (1.14), $(\bar X,\sum_{i=1}^nX_i^2)$ 因而 $(\bar X,S^2)$ 即为充分统计量.

设 $X_1,\cdots,X_n$ 为取自 $N(a,\sigma^2)$ 的 i.i.d. 样本, $Y_1,\cdots,Y_m$ 为取自 $N(b,\sigma^2)$ 的 i.i.d. 样本, 且两样本独立, 参数空间为 $\Theta=\{\theta=(a,b,\sigma^2):a\in\mathbf{R},b\in\mathbf{R},\ \sigma^2>0\}$. 记 $S^2=\big\{\sum_{i=1}^n(X_i-\bar X)^2+\sum_{j=1}^m(Y_j-\bar Y)^2\big\}/(m+n-2)$, 其中 $\bar X,\bar Y$ 分别为各自的样本均值, 则 $(\bar X,\bar Y,S^2)$ 为充分统计量.

设总体分布有密度

$$f_\theta(x)=\frac{cb^{a/c}}{\Gamma(a/c)}x^{a-1}\exp(-bx^c)I(x>0),$$
$$\theta=(a,b,c),\quad a>0,b>0,c>0.$$

设 $X_1,\cdots,X_n$ 为取自上述总体的 i.i.d. 样本, 则无论是转化成指数族, 还是直接应用因子分解定理, 均可得到 $(\prod_{i=1}^n X_i,\sum_{i=1}^n X_i^c)$ 是充分统计量的结论. 上述分布族包括了一些常见分布, 例如, 取 $c=1$ 即得 Gamma 分布, 取 $a=c$ 即得 Weibull 分布.

对于指数族而言, 不论样本量 n 有多大, 在不损失样本信息的前提下, 我们总可以把数据简化为一个 k-维的充分统计量 T, 而 k 是固定的! 这是一些统计问题在指数族中能有圆满解答的关键所在. 在一些条件下已经证明, 若总体分布族有共同支撑, 则只有当它是指数族时, 才有维数与样本量 n 无关的充分统计量. 参看 Barndorff-Nielsen 和 Pedersen (1968), Andersen (1970).

在一些截断分布族中, 虽然不存在共同支撑, 但仍存在维数不随样本量 n 上升的充分统计量.

例 1.10 设 $X_1,\cdots,X_n$ 为从均匀总体 $R(0,\theta),\theta>0$ 中抽取的 i.i.d. 样本, 则由因子分解定理, $T=\max_{1\leqslant i\leqslant n}X_i$ 是充分统计量. 类似地, 如果总体分布为 $R(\theta_1,\theta_2)$, $-\infty<\theta_1<\theta_2<\infty$, 则 $(\min_{1\leqslant i\leqslant n}X_i,\max_{1\leqslant i\leqslant n}X_i)$ 为充分统计量.

例 1.11 设 $X_1,\cdots,X_n$ 为从分布族

$$N(a,\sigma^2),\quad -\infty<a<\infty,\sigma^2\geqslant 0$$

中抽取的 i.i.d. 样本. 证明 $(\bar X,S^2)$ 仍为充分统计量. 此处,

$$\bar X=\frac{1}{n}\sum_{i=1}^n X_i,\quad S^2=\frac{1}{n-1}\sum_{i=1}^n(X_i-\bar X)^2.$$

因为此时分布族已不能为任一 σ-有限的测度所控制, 因而因子分解定理不能直接应用. 记 $X^n=(X_1,\cdots,X_n)$.

当 $-\infty<a<\infty,\sigma^2>0$ 时, $(\bar X,S^2)$ 为充分统计量, 此时对任一 n-维 Borel 集 B, 存在与 (a,σ^2) 无关的条件概率 $P_{(a,\sigma^2)}(X^n\in B|\bar X,S^2)$, 记为 $Q(B|\bar X,S^2)$, 且此时有 $S^2>0$ a.s.; 而当 $\sigma^2=0$ 时, 有 $S^2=0$ a.s..

故当参数空间扩展为 $\{(a,\sigma^2),-\infty<a<\infty,\sigma^2\geqslant 0\}$ 时, 条件概率

$$P_{(a,\sigma^2)}(X^n\in B|\bar X,S^2)=I(S^2>0)Q(B|\bar X,S^2)+I(S^2=0)I_B(X^n)$$

与参数无关. 这就证明了所要的结论.

1.5.3 充分性原则

用统计量 $T=T(X)$代替样本 X, 据以作出统计推断, 相当于把原先的统计模型 I: $\{(\mathcal{X},\mathcal{B}),P_\theta,\theta\in\Theta\}$ 简化为模型 II: $\{(\mathcal{T},\mathcal{G}),P_\theta^T,\theta\in\Theta\}$, 此处 P_θ^T 为由 P_θ 导出的概率分布. 这种简化一般会带来损失. 如前所述, 若用充分统计量 T 代替样本 X, 则所舍弃的并不包含未知分布的任何信息, 因而不会有信息方面的损失. 由于 "信息" 一词并无明确的界定, 这种说法仍需进一步加以明确. 统计决策理论为此提供了一

个精确阐释的平台. 为此, 引进行动空间 $(\mathcal{A}, \mathcal{B}_{\mathcal{A}})$, 损失函数 $L(\theta, a)$, 并对每个决策函数 $\delta(x, A)$, $(x, A) \in \mathcal{X} \times \mathcal{B}_{\mathcal{A}}$, 以 $R(\theta, \delta)$ 记其风险函数. 我们有

定理 1.22 *设充分统计量 T 把统计模型 I 简化为模型 II. 假定存在给定 T 的与 θ 无关的条件概率分布 (参看定理 1.18), 且对每个 θ, 函数 $L(\theta, \cdot)$ 为 $\mathcal{B}_{\mathcal{A}}$-可测, 则对任何决策函数 $\delta(x, A)$, 存在决策函数 $\delta^*(t, A)$, 使*

$$R(\theta, \delta^*) = R(\theta, \delta), \text{ 对 } \forall\, \theta \in \Theta.$$

证明 针对 δ, 在模型 II 之下采用决策函数 δ^* 如下:

$$\delta^*(t, A) = \int_{\mathcal{X}} \delta(x, A) P(dx, t),$$

此处 $P(dx, t)$ 为假定存在的与参数 θ 无关的条件概率分布. 容易证明, δ^* 满足随机化决策函数定义中的两个条件. 依条件分布的定义, 有

$$\int_{\mathcal{T}} P(dx, t) dP_\theta^T(t) = P_\theta(dx) = dP_\theta(x).$$

由 (1.25),

$$R(\theta, \delta) = \int_{\mathcal{X}} \Big\{ \int_{\mathcal{A}} L(\theta, a) \delta(x, da) \Big\} dP_\theta(x).$$

故由 Fubini 定理, 有

$$\begin{aligned} R(\theta, \delta^*) &= \int_{\mathcal{T}} \int_{\mathcal{A}} L(\theta, a) \delta^*(t, da) dP_\theta^T(t) \\ &= \int_{\mathcal{A}} L(\theta, a) \int_{\mathcal{T}} \delta^*(t, da) dP_\theta^T(t) \\ &= \int_{\mathcal{A}} L(\theta, a) \int_{\mathcal{T}} \Big\{ \int_{\mathcal{X}} \delta(x, da) P(dx, t) \Big\} dP_\theta^T(t) \\ &= \int_{\mathcal{A}} L(\theta, a) \int_{\mathcal{X}} \delta(x, da) \int_{\mathcal{T}} P(dx, t) dP_\theta^T(t) \\ &= \int_{\mathcal{A}} L(\theta, a) \int_{\mathcal{X}} \delta(x, da) dP_\theta(x) = R(\theta, \delta). \end{aligned}$$

■

例 1.12 设对某个 σ-有限测度 μ, 样本分布为 $X \sim f(x, \theta) d\mu(x), \theta \in \Theta$. 如所周知, 对给定的样本 x, 函数 $l_x(\cdot) = f(x, \cdot)$ 称为似然函数. 对于给定的 $\theta \in \Theta$, 若定义 $g_\theta(l)$ 为函数 l 在 θ 点的值 $l(\theta)$, 则 $f(x, \theta) = g_\theta(l_x)$. 进一步, 如果一些复杂的可测性条件满足, 则由因子分解定理, 似然函数 $l_x(\cdot)$ 作为一个统计量是充分的. 在此不要有误解, 以为似然函数依赖于参数 θ. 事实上 $l_x(\cdot)$ 作为一个函数完全由样本 x 决定. 因此, 为了对 θ 作出推断或决策, 有关的试验信息全都包含在似然函数 $f(x, \cdot)$ 之中. 进一步, 两个样本的似然函数如果只相差与 θ 无关的一个因子 c_x, 则它们包含了有关 θ 的相同的信息.

考虑到充分性原则, 这可以看作对所谓 "似然原则" 的一个诠释. 似然原则是指, 关于 θ 的推断或决策, 完全基于似然函数. 如果两个样本的似然函数只相差一个因子 c_x, 则应作出同样的推断或决策.

似然原则通常的含义还不止于此, 即使在不同的模型中这个原则仍然适用. 我们在 2.3.6 节讨论 Bayes 统计推断时, 将进一步举例讨论这个问题.

由本节开头关于充分性的解释, 若 T 是充分的且 $T = H(U)$, 则 U 也是充分的. 这是因为 U 的知识蕴含了 T 的知识, 因而可用来对原有数据进行恢复分布的重构. 相对于 U而言, T 对数据提供了更大的简化.

极小充分统计量 称一个充分统计量 T 是极小充分的, 如果对任何充分统计量 U, 存在一个函数 H, 使得 $T = H(U)$ (a. e. $\mathcal{P}$). 可以说, 极小充分统计量对数据提供了最大的简化. 但极小充分统计量并不总是存在的. Bahadur (1957) 证得, **若因子分解定理的条件成立, 且样本空间为欧氏的, 则极小充分统计量存在**. 参看 Lehmann (1983)及所引有关文献. 在一些常见情形, 不难依据因子分解定理找到有关的极小充分统计量.

1.6 分布族的完全性和完全统计量

1.6.1 基本概念, Basu定理

分布族的完全性是数理统计中的基本概念之一, 在最小风险估计的理论中有重要应用. 它是由 Lehmann 和 Scheffé (1950, 1955, 1956) 引进的. 它类似于泛函分析中函数系的完全性的概念, 在那里平方可积函数空间中的一个函数系称为完全的, 如果与该函数系中的每个函数都正交的平方可积函数几乎处处等于 0.

定义 1.9 设 $\{P_\theta, \theta \in \Theta\}$ 为样本空间 $(\mathcal{X}, \mathcal{B})$ 上的一族概率分布. 称分布族 $\{P_\theta, \theta \in \Theta\}$ 是完全的, 若对定义于 $\mathcal{X}$ 上的任何 $\mathcal{B}$-可测函数 f,

$$\mathrm{E}_\theta[f(X)] := \int_{\mathcal{X}} f(x) dP_\theta(x) = 0, \text{ 对 } \forall\, \theta \in \Theta \Longrightarrow\ f(x) = 0 \text{ a. e. } P_\theta, \text{对 } \forall\, \theta \in \Theta.$$

称分布族 $\{P_\theta, \theta \in \Theta\}$ 是有界完全的, 若对定义于 $\mathcal{X}$ 上的任何有界 $\mathcal{B}$-可测函数 f, 上述蕴含关系成立.

称取值于 $(\mathcal{T}, \mathcal{G})$ 的统计量 T 是完全的, 若 $\{P_\theta, \theta \in \Theta\}$ 在 T 之下的导出分布族 $\{P_\theta^T, \theta \in \Theta\}$ 是完全的. 换句话说, 称统计量 T 是完全的, 如果对任何 $\mathcal{G}$-可测函数 g,

$$\mathrm{E}_\theta\big[g\big(T(X)\big)\big] = 0, \text{ 对 } \forall\, \theta \in \Theta \Longrightarrow\ g\big(T(x)\big) = 0 \text{ a.e. } P_\theta, \text{对 } \forall\, \theta \in \Theta.$$

类似地可定义有界完全统计量.

在上述定义中, 如果将参数族 $\{P_\theta, \theta \in \Theta\}$ 代之以一般的分布族 $\mathcal{P}$, 就可得到完全性和完全统计量等的更一般的定义.

容易验证, 分布族 $\{N(a, 1), -\infty < a < \infty\}$ 是完全的, 而分布族 $\{N(0, \sigma^2), \sigma^2 > 0\}$ 是不完全的.

由定义 1.9 可知, 有界完全性的要求弱于完全性, 但要举出一个有界完全而非完全的例子, 也并非轻而易举. 有关的例子见 1.8 节习题 1.46.

关于有界完全性有下面有趣的定理 [参看 Basu (1955) (1958)]. 在一些场合, 它可以提供一些随机变量独立性的简便证明. 详见有关的习题.

定理 1.23 (Basu 定理) 设 X 的样本空间和分布族为 $(\mathcal{X},\mathcal{B})$ 及 $\{P_\theta,\theta\in\Theta\}$, T 为取值于 $(\mathcal{T},\mathcal{G})$ 的有界完全且充分的统计量, V 为取值于 $(\mathcal{V},\mathcal{C})$ 的统计量. 若 V 的分布与 θ 无关, 则 V 与 T 独立.

证明 取 $C\in\mathcal{C}$. 由假定, $a:=P_\theta[V^{-1}(C)]$ 与 θ 无关. 由 T 充分, 知 $\phi(t):=P_\theta[V^{-1}(C)|t]$ 与 θ 无关. 记 $\psi(t)=\phi(t)-a$, 则对任何 $\theta\in\Theta$,

$$\begin{aligned}\mathrm{E}_\theta\{\psi(T)\}&=\mathrm{E}_\theta\{\phi[T(X)]\}-a\\&=\mathrm{E}_\theta\{P_\theta[V^{-1}(C)|T]\}-a\\&=P_\theta[V^{-1}(C)]-a=0.\end{aligned}$$

用 $\{P_\theta^T,\theta\in\Theta\}$ 记 T 的分布族. 由 $\psi(T)$ 有界和 T 的有界完全性, 知

$$\phi(t)=a,\text{a. s. } P_\theta^T,\text{对 }\forall\ \theta\in\Theta.$$

故对任何 $\theta\in\Theta,C\in\mathcal{C},G\in\mathcal{G}$, 有

$$\begin{aligned}P_\theta\big[V^{-1}(C)\cap T^{-1}(G)\big]&=\int_G P_\theta[V^{-1}(C)|t]dP_\theta^T(t)\\&=a\int_G dP_\theta^T(t)=aP_\theta^T(B)\\&=P_\theta\big[V^{-1}(C)\big]P_\theta\big[T^{-1}(G)\big].\end{aligned}$$

这就证明了定理的断言. ■

1.6.2 一些常见分布族的完全统计量

1. 指数型分布族

关于指数族的完全性, 我们有如下的定理.

定理 1.24 设

$$X\sim dP_\theta(x)=C(\theta)\exp\Big\{\sum_{j=1}^k\theta_jT_j(x)\Big\}d\mu(x),$$

其自然参数空间 Θ 在 $\mathbf{R}^k$ 中有内点, 则统计量 $T=\big(T_1(X),\cdots,T_k(X)\big)^\tau$ 是完全的.

证明 设 μ^* 为测度 μ 在 T 之下的导出测度, 则由引理 1.14 (i),

$$T\sim dP_\theta^T(t)=C(\theta)\exp\Big\{\sum_{j=1}^k\theta_jt_j\Big\}d\mu^*(t).$$

其分布族简记为 $\mathcal{P}^T$. 必要时做个参数平移, 不失一般性, 可设 Θ 包含一个矩形 $I=\{(\theta_1,\cdots,\theta_k)^\tau:|\theta_j|<a,j=1,\cdots,k\}$. 设 $f(t)=f^+(t)-f^-(t)$ 满足条件

$$\mathrm{E}_\theta[f(T)]=0,\quad\text{对 }\forall\ \theta\in\Theta,\tag{1.44}$$

其中 $f^+(t)=f(t)\vee0,f^-(t)=[-f(t)]\vee0$. 由定理 1.13, 函数

$$H(\theta):=\int_{\mathcal{T}}f(t)\exp\Big\{\sum_{j=1}^k\theta_jt_j\Big\}d\mu^*(t)$$

在复区域

$$U=\{\theta=(\xi_1+\mathrm{i}\eta_1,\cdots,\xi_k+\mathrm{i}\eta_k)^\tau:(\xi_1,\cdots,\xi_k)^\tau\in I\}$$

上是每个变元的解析函数. 故对固定的实的 $\theta_j\in(-a,a), j=2,\cdots,k$, 将 $H(\theta)$ 视为复变量 $\theta_1=\xi_1+\mathrm{i}\eta_1$ 的函数, 在复区域 $U_1:=\{\theta_1=\xi_1+\mathrm{i}\eta_1:|\xi_1|<a\}$ 上解析, 由 (1.44), 在实轴上的一段恒等于 0, 故在 U_1 上恒等于 0. 再固定 $\theta_1\in U_1$ 以及 $|\theta_j|<a, j=3,\cdots,k$, 视 $H(\theta)$ 为复变量 $\theta_2=\xi_2+\mathrm{i}\eta_2$ 的函数, 在复区域 $U_2:=\{\theta_2=\xi_2+\mathrm{i}\eta_2:|\xi_2|<a\}$ 上解析, 由已证的结果, 在实轴上的一段恒等于 0, 故在 U_2 上恒等于 0. 依此类推, 可知在复区域 U 上, $H(\theta)\equiv 0$. 因而

$$\int_{\mathcal{T}}\exp\Big\{\mathrm{i}\sum_{j=1}^k\eta_jt_j\Big\}f^+(t)d\mu^*(t)\equiv\int_{\mathcal{T}}\exp\Big\{\mathrm{i}\sum_{j=1}^k\eta_jt_j\Big\}f^-(t)d\mu^*(t).\tag{1.45}$$

由此还有 $\int_{\mathcal{T}}f^+(t)d\mu^*(t)=\int_{\mathcal{T}}f^-(t)d\mu^*(t)$, 若其共同值为 0, 则有 $f(t)=0$ a. e. μ^*; 否则不妨假定共同值为 1, 此时 (1.45) 表明, 概率分布 $f^+(t)d\mu^*(t)$ 和 $f^-(t)d\mu^*(t)$ 有相同的特征函数, 因而是相同的分布, 此时也有 $f^+(t)=f^-(t)$ a. e. μ^*, 同样得到 $f(t)=0$ a. e. $\mathcal{P}^T$. 这就是所要证明的. ■

例 1.13 设 $X_1,\cdots,X_n$ i.i.d., 其共同分布为指数分布 $E(a,b), a\in\mathbf{R}, b>0$. 令 $T_1=nX_{(1)}$, $T_2=2\sum_{i=1}^n(X_i-X_{(1)})$, 则两者独立, 且 $T_1\sim E(na,b)$, $T_2\sim b\chi^2_{2n-2}$. 统计量 (T_1,T_2) 既是充分又是完全的. 充分性由因子分解定理得出. 为证完全性, 假定

$$\mathrm{E}_{a,b}[f(T_1,T_2)]=0,\ \text{对}\ \forall\ a,b.$$

记

$$g(t_1,b)=\mathrm{E}_b[f(t_1,T_2)],\tag{1.46}$$

则对任何固定的 $b>0$,

$$\int_s^\infty g(t_1,b)\mathrm{e}^{-t_1/b}dt_1=0,\quad \text{对}\ \forall\ s.$$

对 s 取导数, 知存在 t_1 的一个 Lebesgue 零测集 N_b, 使

$$g(t_1,b)=0,\quad \text{只要}\ t_1\notin N_b.$$

由 Fubini 定理, 对几乎所有的 t_1,

$$g(t_1,b)=0,\quad \text{对几乎所有的}\ b>0.$$

对这些固定的 t_1, 由于 T_2 的分布为一单参数指数族, 由注 1.3 和定理 1.13, $g(t_1,b)=\mathrm{E}_b[f(t_1,T_2)]$ 为 b 的连续函数. 故对几乎所有的 t_1, $\mathrm{E}_b[f(t_1,T_2)]=0$ 对所有的 $b>0$ 成立.

由 T_2 的完全性, 对几乎所有的 t_1, $f(t_1,t_2)=0$ 对 t_2 a.e. 成立. 因而关于二维 Lebesgue 测度, 最终有 $f(t_1,t_2)=0$ a.e.. 这就建立了统计量 (T_1,T_2) 的完全性.

我们在 1.2.3 节中曾经指出, 若 $X\sim Pa(\zeta,\alpha)$, $\zeta>0,\alpha>0$, 则 $Y:=\log X\sim E(\mu,\theta)$, $\mu=\log\zeta\in\mathbf{R}, \theta=1/\alpha>0$. 故若 $X_1,\cdots,X_n$ 为取自上述 Pareto 分布族的

i.i.d. 样本, 则由刚刚证明的结论以及因子分解定理, $(X_{(1)}, \sum_{i=1}^{n} \log X_i)$ 即为 Pareto 分布族的完全充分统计量.

2. 截断分布族

截断分布族有单边截断和双边截断两种, 单边截断又有上下截断之分. 设 (a,b) 为 $\mathbf{R}=(-\infty,\infty)$ 上一个给定的子区间, h 为区间 (a,b)上一个恒正的 Borel 可测函数.

(1) 单边下截断族. 密度函数为

$$f(x,\theta)=h(x)I(\theta<x<b)\Big/\int_{\theta}^{b}h(u)du,\ a<\theta<b,$$

此处要求 $\int_{\theta}^{b}h(u)du<\infty, a<\theta<b.$

(2) 单边上截断族. 密度函数为

$$f(x,\theta)=h(x)I(a<x<\theta)\Big/\int_{a}^{\theta}h(u)du,\ a<\theta<b,$$

此处要求 $\int_{a}^{\theta}h(u)du<\infty, a<\theta<b.$

(3) 双边截断族. 密度函数为

$$f(x,\theta)=h(x)I(\theta_1<x<\theta_2)\Big/\int_{\theta_1}^{\theta_2}h(u)du,\ a<\theta_1<\theta_2<b,$$

此处要求 $\int_{\theta_1}^{\theta_2}h(u)du<\infty, a<\theta_1<\theta_2<b.$

设 $X_1,\cdots,X_n$ 为从上述总体中抽取的 i.i.d. 样本. 记

$$T_1=\min_{1\leqslant i\leqslant n}X_i,\quad T_2=\max_{1\leqslant i\leqslant n}X_i.$$

T_1 和 T_2 分别为前两种情况下的完全充分统计量. 对情形 (3), (T_1,T_2) 为完全充分统计量. 充分性均可由因子分解定理得出. 至于完全性, 前两种情形可由定义直接得出, 情形 (3) 的要复杂一些, 现证明如下.

首先计算 (T_1,T_2) 的密度函数 $f_n(t_1,t_2,\theta)$. 在区间 (a,b) 中任取 $t_1<t_2$, 记 $H(t_1,t_2)=\int_{t_1}^{t_2}h(u)du$, 则

$$\begin{aligned}
&P_\theta(t_i<T_i\leqslant t_i+dt_i,\ i=1,2)\\
&=n(n-1)P_\theta(t_i<X_i\leqslant t_i+dt_i, t_1<X_j\leqslant t_2, i=1,2;\ j=3,\cdots,n)\\
&=n(n-1)h(t_1)dt_1h(t_2)dt_2H^{n-2}(t_1,t_2)\big/H^n(\theta_1,\theta_2),\ \theta_1<t_1<t_2<\theta_2.
\end{aligned}$$

故

$$f_n(t_1,t_2,\theta)=w(t_1,t_2)I(\theta_1<t_1<t_2<\theta_2)\big/H^n(\theta_1,\theta_2),$$

其中 $w(t_1,t_2)=n(n-1)h(t_1)h(t_2)H^{n-2}(t_1,t_2).$

今设统计量 $g(T_1,T_2)$ 满足

$$\mathrm{E}_\theta[g(T_1,T_2)]=0,\ a<\theta_1<\theta_2<b,$$

则

$$\int_{\theta_1<t_1<t_2<\theta_2} g(t_1,t_2)w(t_1,t_2)dt_1dt_2=0,\ a<\theta_1<\theta_2<b. \tag{1.47}$$

上式表明, 如图 1.1 所示, 只要积分区域是在 EH 左上方且斜边落在直线 EH 上的等腰直角三角形, 上述积分即为 0. 在 EH 左上方任取边与坐标轴平行的矩形 $ABCD:=I$, 则由上式,

$$\int_I gw\,dt_1dt_2=\Big\{\int_{\Delta AHE}-\int_{\Delta BGE}-\int_{\Delta DHF}+\int_{\Delta CGF}\Big\}gw\,dt_1dt_2=0.$$

因为这样的 I 是任意的, 且当 $a<t_1<t_2<b$ 时总有 $w(t_1,t_2)>0$, 故关于二维 Lebesgue 测度, 有 $g(t_1,t_2)=0$ a.e.. 这就证明了统计量 (T_1,T_2) 的完全性.

图 1.1

3. 次序统计量的完全性

次序统计量是不是完全的, 取决于是对什么样的分布族而言, 一句话, 就是分布族中必须有足够多的分布. 下面我们介绍陈希孺 (1981, 1997) 提出的一个有力的判别方法. 为此, 设 $X_1,\cdots,X_n$ 为从分布为 $F\in\mathcal{P}$ 的一个一维总体中抽取的 i.i.d. 样本, $T=(X_{(1)},\cdots,X_{(n)})^\tau$ 为它的次序统计量.

定理 1.25 *在分布族 $\mathcal{P}$ 之下次序统计量 T 是完全的, 如果以下两个条件成立:*

(i) $F\in\mathcal{P},G\in\mathcal{P},0<p<1\Longrightarrow\ pF+(1-p)G\in\mathcal{P}$.

(ii) $F\in\mathcal{P},S=(a,b),$ *而* $F(S)>0\Longrightarrow\ F_S\in\mathcal{P}$, *此处, 分布* F_S *定义为: 对任一* Borel *集* A, $F_S(A)=F(A\cap S)/F(S)$.

常见的一维非参数分布族大多满足这两个条件, 但参数族则不然. 证明和举例都可参看陈希孺 (1981, 1997). 在该书中, 他还把这一定理推广到多样本情形.

定理 1.26 设 $X_{i1},\cdots,X_{in_i}$ 为从分布为 $F_i\in\mathcal{P}_i$ 的一维总体中抽取的 i.i.d. 样本, $i=1,\cdots,c$, 且这 c 个样本全体独立. 第 i 个样本的次序统计量记为 $(X_{i(1)},\cdots,X_{i(n_i)})$. 令

$$T=(X_{1(1)},\cdots,X_{1(n_1)};\cdots;X_{c(1)},\cdots,X_{c(n_c)}).$$

如果每个 $\mathcal{P}_i$ 都满足定理 1.25 的条件 (i)、条件 (ii), 则 T 为完全统计量.

1.7 凸损失函数

由于凸分析在统计决策问题中的重要性, 特别是在点估计理论中的突出地位, 为便于后面的应用, 在此我们专门对凸损失之下风险理论的一些重要结果做一介绍.

在点估计的决策理论中, 人们经常使用的是所谓的凸损失. 一个损失函数称为凸的, 如果对每个固定的 $\theta\in\Theta$, $L(\theta,\cdot)$ 为 $\mathcal{A}$ 上的一个凸函数, 此时 $\mathcal{A}$ 为 $\mathbf{R}^m$ 或是一个 m-维的开凸集. 形如 $L(\theta,a)=||g(\theta)-a||^2$ 的平方损失是凸损失中最重要的例子. 在凸损失之下有关 $g(\theta)$ 的估计问题, 我们可以得出一些什么样的简洁而重要的结果呢? 要回答这个问题, 首先要引进下述著名不等式.

定理 1.27 (Jensen 不等式) 设 X 为取值于一个 m-维开凸集 C 的随机向量, f 为 C 上的一个凸函数, 且 $\mathrm{E}(X)$ 存在有限, 则

$$\mathrm{E}[f(X)]\geqslant f[\mathrm{E}(X)].$$

当 f 为严格凸时, 上式中的等号当且仅当 $P(X=\mathrm{E}X)=1$ 时成立.

证明 不失一般性, 可设 $C=\mathbf{R}^m$. 由凸分析, 过凸函数 $y=f(x)$ 图像上的任一点, 可在 $\mathbf{R}^{m+1}$ 中作一支撑平面, 使图像在平面上方. 以 $y=f(\mathrm{E}(X))+\alpha^\tau(x-\mathrm{E}(X))$ 记图像在点 $(\mathrm{E}(X),f(\mathrm{E}(X)))$ 处的一个支撑平面, 则

$$f(x)\geqslant f(\mathrm{E}(X))+\alpha^\tau(x-\mathrm{E}(X)),\quad x\in\mathbf{R}^m,$$

因此

$$\mathrm{E}[f(X)]\geqslant f(\mathrm{E}(X))+\alpha^\tau\mathrm{E}[X-\mathrm{E}(X)]=f(\mathrm{E}(X)).\tag{1.48}$$

若 f 为严格凸, 则

$$f(x)>f(\mathrm{E}(X))+\alpha^\tau(x-\mathrm{E}(X)),\ \text{只要 } x\neq\mathrm{E}(X),$$

因而只有在 $P(X=\mathrm{E}(X))=1$ 时, (1.48) 中的等号才能成立. ■

注意这个定理并不排除 $\mathrm{E}[f(X)]=\infty$ 的可能性.

例 1.14 设 μ 为样本空间 $(\mathcal{X},\mathcal{B})$ 上的一个 σ 有限测度, $f(x)d\mu(x)$ 和 $g(x)d\mu(x)$ 为两个概率分布. 密度函数 f 与 g 之间相对于 f 的熵距离 (entropy distance) 可以定义为

$$\mathrm{E}_f\log[f(X)/g(X)]:=\int\log[f(x)/g(x)]f(x)d\mu(x).$$

它也可称为 g 在 f 处的 Kullback-Leibler 信息量或 g 和 f 之间的 Kullback-Leibler 距离. 由 Jensen 不等式, 我们立得下面的引理.

引理 1.28 对于任意的概率密度函数 f 与 g,

$$\mathrm{E}_f \log[f(X)/g(X)] = -\mathrm{E}_f \log[g(X)/f(X)] \geqslant -\log[\mathrm{E}_f(g(X)/f(X))] = 0. \tag{1.49}$$

等式成立当且仅当 $f = g$. 又若

$$\mathrm{E}_f|\log f(X)| := \int |\log f(x)| f(x) d\mu(x) < \infty,$$

则

$$\mathrm{E}_f \log[g(X)] \leqslant \mathrm{E}_f \log[f(X)].$$

这是一个很有用的不等式.

引理 1.29 设 $\rho(\cdot)$ 是 $\mathbf{R}^m$ 上的一个凸函数, 且 $\lim_{||u||\to\infty} \rho(u) = \infty$, 则存在 $c > 0$ 和 c_1, 使得

$$\rho(u) \geqslant c||u|| + c_1.$$

证明 因为 $\lim_{||u||\to\infty} \rho(u) = \infty$, 故存在 r_0 充分大, 使 $c = \inf_{||u||=r_0} \rho(u) - \rho(0) > 0$. 为简化记号, 不妨设 $r_0 = 1$. 因为凸函数 ρ 为连续函数, 故 $h := \min_{||u||\leqslant 1} \rho(u)$ 存在. 单位球面外任一点均可写成 $r\alpha$ 的形式, 其中 α 在球面上, 而 $r > 1$. 由 ρ 的凸性, $\rho(\alpha) \leqslant \frac{r-1}{r}\rho(0) + \frac{1}{r}\rho(r\alpha)$, 即

$$\rho(r\alpha) \geqslant r[\rho(\alpha) - \rho(0)] + \rho(0) \geqslant cr + \rho(0).$$

而当 $0 \leqslant r \leqslant 1$ 时, $\rho(r\alpha) \geqslant h \geqslant cr + (h - c)$.

由最后两个不等式即可得到所要证明的结果. ■

定理 1.30 设样本空间 $(\mathcal{X}, \mathcal{B})$ 和行动空间 $(\mathcal{A}, \mathcal{B}_\mathcal{A})$ 均为欧氏的, $\mathcal{A} = \mathbf{R}^m$, $X \sim P_\theta : \theta \in \Theta$, 对 $\forall\, \theta, L(\theta, \cdot)$ 均为 $\mathcal{A}$ 上的凸函数, 且

$$\lim_{||a||\to\infty} L(\theta, a) = \infty, \quad \theta \in \Theta.$$

则非随机化决策函数的类构成一个本质完全类.

证明 设 $\delta(x, da)$ 为一随机化决策函数. 记

$$B = \left\{ x \in \mathcal{X} : \int_\mathcal{A} ||a|| \delta(x, da) < \infty \right\}.$$

定义

$$\delta_0(x) = \begin{cases} \int_\mathcal{A} a\delta(x, da), & \text{若 } x \in B, \\ 0, & \text{若 } x \notin B. \end{cases}$$

当 $x \in B$ 时, 由凸函数的 Jensen 不等式,

$$L(\theta, \delta_0(x)) = L\left(\theta, \int_\mathcal{A} a\delta(x, da)\right) \leqslant \int_\mathcal{A} L(\theta, a)\delta(x, da).$$

由引理 1.29, 存在 $c(\theta) > 0$ 和 $c_1(\theta)$, 使 $L(\theta, a) \geqslant c(\theta)||a|| + c_1(\theta)$. 故当 $x \notin B$ 时,

$$\int_\mathcal{A} L(\theta, a)\delta(x, da) \geqslant \int_\mathcal{A} [c(\theta)||a|| + c_1(\theta)]\delta(x, da) = \infty.$$

由以上两式, 对所有 x, $L(\theta, \delta_0(x)) \leqslant \int_\mathcal{A} L(\theta, a)\delta(x, da)$. 故 $R(\theta, \delta_0) \leqslant R(\theta, \delta), \theta \in \Theta$. 这就证明了所要的结论. ■

参看成平, 陈希孺, 陈桂景, 吴传义 (1985), 第 52 页.

定理 1.31 (Rao-Blackwell 定理) 设 X 的样本分布族为 $\mathcal{P}=\{P_\theta:\theta\in\Theta\}$, 要估计参数函数 $g(\theta)$, 采用的损失函数 $L(\theta,a)$ 为 a 的凸函数. 又设 T 关于 $\mathcal{P}$ 是充分的, $\delta(X)$ 是 $g(\theta)$ 的一个估计, 对任一 P_θ 都有有限的期望, 则可构造这样一个估计 $\eta(T)=\mathrm{E}[\delta(X)|T]$, 使其风险满足

$$R(\theta,\eta(T))\leqslant R(\theta,\delta),\ \theta\in\Theta.$$

若 L 对 a 为严格凸, 且对某个 θ, 上式等号成立, 而 $R(\theta,\delta)=\mathrm{E}_\theta[L(\theta,\delta(X))]<\infty$, 则 $\delta(X)=\eta[T(X)]$ a.s. P_θ.

在这个定理中, 并不要求样本空间 $(\mathcal{X},\mathcal{B})$ 为欧氏的.

证明 因为统计量 T 关于分布族 $\{P_\theta,\ \theta\in\Theta\}$是充分的, 且 $\delta(X)$ 具有有限的期望, 故对此 $\delta(X)$, 存在条件期望 $\mathrm{E}_\theta[\delta(X)|T]$ 的一个版本与 θ 无关 (参看 1.8 节习题 1.23), 记之为 $\eta(T):=\mathrm{E}[\delta(X)|T]$. 由假定, $\mathrm{E}_\theta|\delta(X)|<\infty,\theta\in\Theta$, 故由条件 Jensen 不等式,

$$\mathrm{E}_\theta|\eta(T)|=\mathrm{E}_\theta\big|\mathrm{E}[\delta(X)|T]\big|\leqslant\mathrm{E}_\theta\{\mathrm{E}[|\delta(X)|\big|T]\}=\mathrm{E}_\theta|\delta(X)|<\infty,\quad\theta\in\Theta,$$

故对任一 P_θ, $\eta(T):=\mathrm{E}[\delta(X)|T]$ a.s. 有限, 又因为其与参数无关, 故可作为 $g(\theta)$ 的一个估计. 由条件 Jensen 不等式,

$$L(\theta,\eta(T))=L(\theta,\mathrm{E}[\delta(X)|T])\leqslant\mathrm{E}_\theta[L(\theta,\delta(X))|T],\tag{1.50}$$

因而对任一 $\theta\in\Theta$,

$$\mathrm{E}_\theta[L(\theta,\eta(T))]\leqslant\mathrm{E}_\theta\{\mathrm{E}_\theta[L(\theta,\delta(X))|T]\}=\mathrm{E}_\theta[L(\theta,\delta(X))].\tag{1.51}$$

设 $R(\theta,\delta)=\mathrm{E}_\theta[L(\theta,\delta(X))]<\infty$, 则由 (1.51) 中的最后一个等式, 有

$$\mathrm{E}_\theta[L(\theta,\delta(X))|T]<\infty,\quad\text{a.s. }P_\theta.\tag{1.52}$$

今若 L 为严格凸, 且对该 θ, (1.51) 中等号成立, 考虑到 (1.52), 则 (1.50) 中等号以概率 1 成立, 由定理 1.27, $P_\theta\{\delta(X)=\eta[T(X)]\}=1$. ■

这一定理最初是 Rao (1945) 和 Blackwell (1947) 在平方损失下建立的, Hodges 和 Lehmann (1950) 把它推广到一般的凸损失.

这个定理也适用于一般的统计决策问题. 由以上两个定理, 立得如下推论.

推论 1.32 在定理 1.30 的条件下, 基于充分统计量的非随机化决策函数的类构成一个本质完全类.

1.8 习题和补充

在以下陈述中, $I(A)$ 表示集合 A 的示性函数. 设 $(\Omega,\mathcal{A},P)$ 为一概率空间, $\mathcal{C}\subset\mathcal{A}$, 则由 $\mathcal{C}$ 生成的子 σ-域常简记为 $\sigma(\mathcal{C})$.

1.1 证明二元分布函数 $F(x,y)$ 为 Borel 可测函数.

1.2 设 $\mathrm{E}(X)=\mathrm{E}(Y)=0,\mathrm{E}(X^2)=\mathrm{E}(Y^2)=1,\mathrm{E}(XY)=\rho$. 证明: 对任一 $a>0$,

$$P\{\max\{|X|,|Y|\}\geqslant a\}\leqslant(1+\sqrt{1-\rho^2})/a^2\ \text{ (二元 Chebyshev 不等式).}$$

[证法一: 不妨设 $\rho \geqslant 0$. 对 $\forall\, t \in [0,1), \forall\, x, y$, $x^2 - 2txy + y^2 \geqslant (1-t^2)(x^2 \vee y^2)$. 知 $\{\max\{|X|, |Y|\} \geqslant a\} \subset \{X^2 - 2tXY + Y^2 \geqslant (1-t^2)a^2\}$. 用 Chebyshev 不等式, 并选取 t 最小化不等式右方.

证法二 (更为简单): 令 $U = (X^2+Y^2)/2, V = (X^2-Y^2)/2$, 得 $\max\{X^2, Y^2\} = U + |V|$. 相继用 Markov 及 Schwarz 不等式.]

1.3 设 X, Y 独立同分布, 证明: $\mathrm{E}|X+Y| \geqslant \mathrm{E}|X|$.

[提示: 不妨设 $\mathrm{E}|X| < \infty$, 且设 $X \sim F$, 则 $-X \sim G(x) = 1 - F(-x-0)$.

证法一: 只需证明 $\mathrm{E}(|X+Y|-|X-Y|) = 2\int_0^\infty [1-F(-t)-F(t)]^2 dt \geqslant 0$. 为此, 利用 $\mathrm{E}(|X+Y|-|X-Y|)I(Y>0) = \mathrm{E}\int I(0<t<Y)[\mathrm{sgn}(X+t)+\mathrm{sgn}(X-t)]dt$. 为得到 $\mathrm{E}(|X+Y|-|X-Y|)I(Y<0)$ 的积分表达式, 只需将 X, Y 换成 $-X, -Y$.

证法二: 用 $\mathrm{E}(|X+Y|-|X|)I(Y>0) = \mathrm{E}\int I(0<t<Y)\mathrm{sgn}(X+t)dt$, 以及 $\mathrm{E}(\cdots)I(Y<0)$ 的类似表达式, 可将 $\mathrm{E}(|X+Y|-|X|)$ 化为关于 dt 的一个积分, 继而证明被积函数非负.]

1.4 设 X_1, X_2 独立同分布, $X_1 \sim R(1,\theta), \theta > 1$. 求 $Y = X_1X_2$ 的密度函数.

1.5 设 X_1, X_2, X_3 为 i.i.d. $N(0,1)$ 变量, $U = X_1 + X_2 + X_3, V = X_1 - X_2, W = X_1 + X_2 - 2X_3$. 求 $Z = \sqrt{2/3}|U/V|$ 和 $T = (2U^2+3V^2)/(2U^2+3V^2+W^2)$ 的密度函数.

1.6 设 $(X,Y) \sim N(0,0;1,1,\rho)$, 证明: $\mathrm{E}\big[\max\{X,Y\}\big] = \sqrt{(1-\rho)/\pi}$.

[提示: 令 $U = X+Y, V = X-Y$.]

*1.7 设 $\theta \sim R(0,2\pi), \psi = \mathrm{arccot}[(\cos\theta + h)/\sin\theta], 0 < h \leqslant 1$. 证明: $\psi \sim R(0,\pi)$.

1.8 设 X, Y 为 i.i.d. $N(0,1)$ 随机变量. 求 $Z = Y/X$ 的密度函数.

1.9 设 X 为期望有限的一维随机变量, 其分布关于零点对称, a 为一正数. 令 $Y = XI(|X| \leqslant a)$, $Z = XI(|X| > a)$. 证明: $\mathrm{E}(Z|Y) = 0$ a.s..

1.10 假定有关的期望都有限, 证明:

$$\mathrm{E}[X\mathrm{E}(Y|Z)] = \mathrm{E}[Y\mathrm{E}(X|Z)], \quad \mathrm{E}[\mathrm{E}(Z|X,Y)|X] = \mathrm{E}(Z|X),$$
$$\mathrm{Var}\,(X) = \mathrm{E}[\mathrm{Var}\,(X|T)] + \mathrm{Var}\,[\mathrm{E}(X|T)],$$

对于 $\mathrm{Cov}(X,Y)$ 也有类似的等式.

1.11 证明条件 Jensen 不等式: 设 φ 为定义于 d-维开凸集 K 上的一个凸函数, X 为取值于 K 的可积随机向量, T 为一统计量, 且 $\mathrm{E}|\varphi(X)| < \infty$, 则

$$\varphi(\mathrm{E}[X|T]) \leqslant \mathrm{E}[\varphi(X)|T].$$

如果 φ 在 K 上是严格凸的, 则上式中等号成立当且仅当 $X = \mathrm{E}[X|T]$ a.s..

1.12 (i) 设 X, Y 均为一维随机变量, $\mathrm{E}(X^2) < \infty$, 且 $\mathrm{E}[X|Y] = Y$ a.s., $\mathrm{E}[Y|X] = X$ a.s.. 证明: $X = Y$ a.s..

(ii) 证明: 若仅要求 $\mathrm{E}(|X|) < \infty$, 且 $\mathrm{E}[X|Y] = Y$ a.s., $\mathrm{E}[Y|X] = X$ a.s., 则仍有 $X = Y$ a.s..

1.13 设 $X_1,\cdots,X_n$ i.i.d., $P_\theta(X_1=1)=1-P_\theta(X_1=0)=\theta, 0<\theta<1$. 记 $T=\sum_{i=1}^n X_i$. 求 $P_\theta\big(X_1(1-X_2)|T=t\big), n\geqslant 2$.

1.14 设随机向量 (X,Y) 服从二元正态分布 $N(a,b;\sigma_1^2,\sigma_2^2,\rho)$. 求 Y 在给定 $X=x$ 下的条件分布.

1.15 设 $X_1,\cdots,X_n$ i.i.d., X_1 服从正态分布 $N(0,\sigma^2)$, $T=(X_1^2+\cdots+X_n^2)^{1/2}$. 证明: 在给定 $T=t$ 时 $(X_1,\cdots,X_n)$ 的条件分布由下式给出:

$$P(A,t)=|S_t\cap A|/|S_t|.$$

此处, $A\in\mathcal{B}_n$, S_t 为 $\mathbf{R}^n$ 中以原点为中心、t 为半径的超球面, $|S_t\cap A|$ 为该超球面上的区域 $S_t\cap A$ 的面积.

1.16 设 $X_1,\cdots,X_n$ i.i.d., X_1 有连续的分布函数 F, $X_{(1)}\leqslant X_{(2)}\leqslant\cdots\leqslant X_{(n)}$ 为它的次序统计量. 定义

$$\bar{F}(x)=\begin{cases}[F(x)-F(x_{(1)})]/[1-F(x_{(1)})], & \text{当 } x>x_{(1)},\\ 0, & \text{当 } x\leqslant x_{(1)},\end{cases}$$

则在 $X_{(1)}=x_{(1)}$ 给定的条件下, $X_{(2)},\cdots,X_{(n)}$ 的条件分布即为从截断分布 $\bar{F}$ 中抽取的容量为 $n-1$ 的次序样本的分布.

1.17 设 $(\Omega,\mathcal{A},P)$ 为一概率空间, $\Omega=[0,1]$, $\mathcal{A}$ 为 $[0,\ 1]$ 区间上的 Borel σ-域, P 为集合的 Lebesgue 测度. 今设 $\mathcal{A}_0=\sigma\{[0,1/3),[1/3,2/3),[2/3,1]\}$, $X_1(\omega)=\omega^2, X_2(\omega)=-\omega,\omega\in[0,1]$, 求 $\mathrm{E}(X_i|\mathcal{A}_0), i=1,2$.

1.18 设概率空间跟 1.17 题一样, $\mathcal{A}_0=\sigma\{[1/2-a,\ 1/2+a], 0\leqslant a\leqslant 1/2\}$, $X_1(\omega)=\omega$, $X_2(\omega)=2\omega^2-\omega,\omega\in[0,1]$, 求 $\mathrm{E}(X_i|\mathcal{A}_0), i=1,2$.

1.19 设 $X_1,\cdots,X_n$ i.i.d., $X_1\sim E(0,1)$, $X_{(1)}\leqslant X_{(2)}\leqslant\cdots\leqslant X_{(n)}$ 为它的次序统计量, 则在 $X_{(1)}$ 给定的条件下, $\big(X_{(2)}-X_{(1)},\cdots,X_{(n)}-X_{(1)}\big)$ 的条件分布等同于从分布为 $E(0,1)$ 的总体中抽取的容量为 $n-1$ 的次序样本的分布.

1.20 设 $X_1,\cdots,X_n$ i.i.d., X_1 有分布 $f(x)dx$, $X_{(1)}=\min\{X_1,\cdots,X_n\}$ 而 $g(X_1)$ 可积. 证明:

$$\mathrm{E}[g(X_1)|x_{(1)}]=\frac{1}{n}g(x_{(1)})+\frac{n-1}{n}\int_{x_{(1)}}^{\infty}g(u)f(u)du\Big/\int_{x_{(1)}}^{\infty}f(u)du.$$

1.21 设 X,Y i.i.d., 其共同分布为指数分布 $\mathrm{e}^{-x}I(x>0)dx$. 设 u 和 s 均为正数. 试求 $P(X>u|X\wedge Y=s)$ 和 $\mathrm{E}(X|X\wedge Y)$.

1.22 设 X 的分布同上. 对每一 $t>0$ 计算 $\mathrm{E}[X|X\wedge t]$ 和 $\mathrm{E}[X|X\vee t]$.

1.23 设统计量 T 关于分布族 $\{P_\theta,\theta\in\Theta\}$ 是充分的, 函数 f 对每个 $\theta\in\Theta$ 是 $(\mathcal{B},P_\theta)$-可积的, 则存在条件期望函数 $\mathrm{E}_\theta[f(X)|t]$ 的一个版本与 θ 无关.

1.24 设 $X_1,\cdots,X_n$ 为 i.i.d. 指数 $E(0,1)$ 变量, $(X_{1:n},\cdots,X_{n:n})$ 为其次序统计量, 则 $nX_{1:n},\cdots,(n-r+1)(X_{r:n}-X_{r-1:n}),\cdots,X_{n:n}-X_{n-1:n}$ 仍为 i.i.d. 指数 $E(0,1)$ 变量.

1.25 设 $X_1,\cdots,X_n$ 为 i.i.d. 指数 $E(0,1)$ 变量, $(X_{1:n},\cdots,X_{n:n})$ 为其次序统计量,

$Y_r=\sum_{i=1}^r X_i/(n-i+1), r=1,\cdots,n$, 则

$$(X_{1:n},\cdots,X_{n:n})\overset{d.}{=}(Y_1,\cdots,Y_n).$$

1.26 设 $X_1,\cdots,X_n$ i.i.d., 其共同分布为指数分布 $E(a,b),-\infty<a<\infty,b>0$. 令 $T_1:=nX_{(1)}$, $T_2:=\sum_{i=2}^n(X_{(i)}-X_{(1)})$. 证明 T_1,T_2 独立, 且 $T_1\sim E(na,b)$, $2T_2/b\sim\chi^2_{2n-2}$.

1.27 设 $X_1,\cdots,X_n$ i.i.d., $X_1\sim E(a,b),a\in\mathbf{R},b>0$, $X_{(1)}\leqslant X_{(2)}\leqslant\cdots\leqslant X_{(n)}$ 为它的次序统计量. 假定我们只能观测到 $\left(X_{(1)},X_{(2)},\cdots,X_{(r)}\right)$, 其中 $r\geqslant 2$ 是事先指定的一个整数. 这就是指数寿命数据的 II 型删失模型, 即定数删失模型. 令 $T_1:=nX_{(1)}$, $T_2:=\sum_{i=2}^r\left(X_{(i)}-X_{(1)}\right)+(n-r)X_{(r)}$. 证明 T_1,T_2 独立, 且 $T_1\sim E(na,b)$, $2T_2/b\sim\chi^2_{2r-2}$.

1.28 设 $X_1,\cdots,X_n$ i.i.d., $X_1\sim E(0,\theta),\theta>0$, $X_{(1)}\leqslant X_{(2)}\leqslant\cdots\leqslant X_{(n)}$ 为它的次序统计量. 给定整数 $r\geqslant 2$, 假定我们只能观测到 $\left(X_{(1)},X_{(2)},\cdots,X_{(r)}\right)$. 令 $T:=\sum_{i=1}^r X_{(i)}+(n-r)X_{(r)}$. 证明 $2T/\theta\sim\chi^2_{2r}$.

1.29 设 $U_{1:n},\cdots,U_{n:n}$ 为取自 $U(0,1)$ 的次序样本, $X_1,\cdots,X_n$ 为 i.i.d. 指数 $E(0,1)$ 变量, $Y_r=\sum_{i=1}^{n-r+1}Z_i/(n-i+1), r=1,\cdots,n$, 则

$$(-\log U_{1:n},\cdots,-\log U_{n:n})\overset{d.}{=}(Y_1,\cdots,Y_n).$$

1.30 设 $(U_{1:n},\cdots,U_{n:n})$ 为取自 $U(0,1)$ 的次序样本, $X_1,\cdots,X_{n+1}$ 为 i.i.d. 指数 $E(0,1)$ 变量, $S_r=\sum_{i=1}^r X_i, r=1,\cdots,n+1$, 则

$$(U_{1:n},U_{2:n}-U_{1:n},\cdots,U_{n:n}-U_{n-1:n},1-U_{n:n})\overset{d.}{=}(X_r/S_{n+1},r=1,\cdots,n+1).$$

一个等价的命题为:

$$(U_{1:n},\cdots,U_{n:n})\overset{d.}{=}(S_1/S_{n+1},\cdots,S_n/S_{n+1}).$$

1.31 设 $(V_0,V_1,\cdots,V_r)\sim D(r;\ a_0,a_1,\cdots,a_r)$, $a_0>0,\cdots,a_r>0$, $r\geqslant 3$, 则由 Dirichlet 分布的定义, 无需计算即可证明:

(1) $(V_0+V_1,V_2,\cdots,V_r)\sim D(r-1;\ a_0+a_1,a_2,\cdots,a_r)$.

(2) $V_1/(V_1+V_2)\sim Be(a_1,a_2)$.

1.32 设 $(X,Y)\sim N(0,0;\sigma_1^2,\sigma_2^2,\rho),\sigma_1>0,\sigma_2>0,|\rho|<1$. 证明:

$$\frac{1}{1-\rho^2}\left(\frac{X^2}{\sigma_1^2}-\frac{2\rho XY}{\sigma_1\sigma_2}+\frac{Y^2}{\sigma_2^2}\right)\sim\chi^2_2.$$

1.33 设 $X=(X_1,\cdots,X_n)^\tau\sim N_n(\mu\mathbb{1}_n,\Sigma)$, 其中 $\mathbb{1}_n=(1,\cdots,1)^\tau$

$$\Sigma=\sigma^2\begin{pmatrix}1&\rho&\cdots&\rho\\ \rho&1&\cdots&\rho\\ \vdots&\vdots&&\vdots\\ \rho&\rho&\cdots&1\end{pmatrix},$$

$|\rho|<1$. 证明 $\bar X$ 和 $W=\sum_{i=1}^n(X_i-\bar X)^2$ 独立, $\bar X\sim N(\mu,\ [1+(n-1)\rho]\sigma^2/n)$,

$W/[(1-\rho)\sigma^2] \sim \chi^2_{n-1}$.

[提示: 作正交变换 $Y = CX$, 其中 $C^\tau = (\mathbb{1}_n/\sqrt{n}, C_1)$ 为一正交阵.]

1.34 设 $X_1, \cdots, X_n$ 为 i.i.d. 正态变量, 记 $\bar{X} = \sum_{i=1}^n X_i/n, S^2 = \sum_{i=1}^n (X_i - \bar{X})^2$, 则 $\sqrt{n-1}(X_1 - \bar{X})/S \overset{d.}{=} (n-1)Y/\sqrt{n(Y^2+n-2)}$, 其中 $Y \sim t_{n-2}$.

[提示: 作 $(X_1, \cdots, X_n)^\tau$ 到 $(Z_1, \cdots, Z_n)^\tau$ 的正交变换, 使 $Z_1 = \sqrt{n}\bar{X}, Z_2 = \sqrt{n/(n-1)}(X_1 - \bar{X})$.]

1.35 设 $(X_1, Y_1), \cdots, (X_n, Y_n)$ 是从二维正态总体 $N(\theta_1, \theta_2; 1, 1, \rho)$ 中抽取的 i.i.d. 样本. 记 $S_X^2 = \sum_{i=1}^n (X_i - \bar{X})^2, S_Y^2 = \sum_{i=1}^n (Y_i - \bar{Y})^2$. 用下面的方法可求得样本相关系数

$$r = \sum_{i=1}^n (Y_i - \bar{Y})X_i / S_X S_Y$$

的分布. 设 $\rho = 0$, 给定 $Y_1, \cdots, Y_n$, 则 $(X_1, \cdots, X_n)$ 的条件分布等同于无条件分布. 此时可用与 1.34 题类似的正交变换. 这种方法也适用于 $\rho \neq 0$ 的情形, 但计算要复杂得多.

*1.36 设 $T_n \sim t_n$, 其密度函数为 $f_n(t) = f_n(0)(1 + x^2/n)^{-(n+1)/2}$.

(1) 证明: $f_n(0) > f_{n-1}(0)$.

(2) 证明: 方程 $f_n(x) = f_{n-1}(x)$ 在 $[0, \infty)$ 内有且只有唯一解.

(3) 基于上面的结论, 证明: 对任一 $t > 0$, $P(|T_n| > t) < P(|T_{n-1}| > t)$. 一个等价的命题是: 若 $0 < \alpha < 1/2$, 则分布 t_n 的上 α 分位数 $t_n(\alpha)$ 随着 n 的上升而下降.

1.37 设 X_1, X_2 独立, $X_1 \sim \Gamma(\kappa_1, \beta), X_2 \sim \Gamma(\kappa_2, \beta), \kappa_1 > 0, \kappa_2 > 0, \beta > 0$. 求 $Z = (X_2/\kappa_2)/(X_1/\kappa_1)$ 的密度函数.

1.38 设 $X_1, \cdots, X_n$ 相互独立, $X_i \sim G(\lambda_i, 1), \lambda_i > 0, i = 1, \cdots, n$. 令

$$Y_1 = \sum_{i=1}^n X_i, \quad Y_r = X_r \Big/ \sum_{i=1}^{r-1} X_i, \ r = 2, \cdots, n.$$

证明 $Y_1, \cdots, Y_n$ 也是独立变量, 同时给出它们各自的分布.

1.39 设 X_1, X_2, X_3 为 i.i.d. $N(0, \sigma^2)$ 变量.

(1) 证明: $(X_1 + X_2X_3)/\sqrt{1 + X_2^2} \sim N(0, \sigma^2)$.

(2) 证明: X_1/X_2 与 $X_1^2 + X_2^2 + X_3^2$ 独立.

1.40 设 $X_1, \cdots, X_n$ i.i.d., $X_1 \sim E(0, \theta), \theta > 0$, 记 $T = \sum_{i=1}^n X_i$.

(1) 求 X_1/T 的概率分布.

(2) 求 X_1 在给定 $T = t$ 时的条件分布 $(t > 0)$.

1.41 设 $X_1, \cdots, X_n$ 为取自正态总体 $N(\theta, \theta^2)$ 的 i.i.d. 样本, $\theta > 0$ 是未知参数. 证明统计量 $(\sum_{i=1}^n X_i, \sum_{i=1}^n X_i^2)$ 是充分的但不是有界完全的.

1.42 设 $X_1, \cdots, X_m$ 和 $Y_1, \cdots, Y_n$ 分别是来自 $N(a, \sigma^2)$ 和 $N(a, \rho\sigma^2)$ 的 i.i.d. 样本,

两样本整体独立, $\rho > 0$ 已知, $a \in (-\infty, \infty)$ 和 $\sigma > 0$ 未知. 试给出两样本联合分布族的一个完全充分统计量.

1.43 设 $X_1, \cdots, X_n$ i.i.d., $X_1 \sim R(\theta - 1/2, \theta + 1/2), \theta \in \mathbf{R}^1$. 证明 $(\min_{1 \leqslant i \leqslant n} X_i, \max_{1 \leqslant i \leqslant n} X_i)$ 为一充分统计量, 但不是完全的.

1.44 设 $X_1, \cdots, X_n$ 和 $Y_1, \cdots, Y_n$ 分别为取自正态总体 $N(\mu, \sigma_1^2)$ 和 $N(\mu, \sigma_2^2)$ 的 i.i.d. 样本, 且两样本独立, 此处 $\mu \in (-\infty, \infty), \sigma_1 > 0, \sigma_2 > 0$. 证明充分统计量 $(\sum_{i=1}^n X_i, \sum_{i=1}^n Y_i, \sum_{i=1}^n X_i^2, \sum_{i=1}^n Y_i^2)$ 不是有界完全的.

1.45 证明分布族 $\{\chi^2_{n,\delta} : \delta \geqslant 0\}$ 和 Beta 分布族 $\{Be(a,b), a > 0, b > 0\}$ 都是完全分布族.

1.46 设 $\mathcal{X} = \{-1, 0, 1, 2, \cdots\}$,

$$P_\theta(X = -1) = \theta, \quad P_\theta(X = x) = (1-\theta)^2\theta^x, \quad x = 0, 1, 2, \cdots.$$

证明分布族 $\{P_\theta, 0 < \theta < 1\}$ 是有界完全而不是完全的.

1.47 设 $X_1, \cdots, X_n$ 为从均匀分布 $U(-\theta, \theta)$ 中抽取的 i.i.d. 样本, $\theta > 0$. 记 $M = \max_{1 \leqslant i \leqslant n} X_i, m = \min_{1 \leqslant i \leqslant n} X_i, T_1 = M - m, T_2 = \max\{M, -m\}$.

(1) 证明 T_2 是充分统计量, 且 T_1/T_2 与 T_2 独立.

(2) 求 $T_1/(2T_2)$ 的分布.

*1.48 设 $X = (X_1, \cdots, X_p)^\tau \sim N(0, I_p)$, 此处 I_p 为 p 阶单位方阵. 则

(1) $X \overset{d.}{=} RU$, 其中, $R^2 \sim \chi^2_p$, U 服从 p-维空间单位球面 S_{p-1} 上的均匀分布 $U(S_{p-1})$.

(2) $X \overset{d.}{=} R^*V$, 其中, $R^{*2} \sim \chi^2_{p+2}$, V 服从 p-维单位球体 V_p 上的均匀分布 $U(V_p)$.

[提示: p-维单位球体 V_p 的体积和表面积分别为 $\pi^{p/2}/\Gamma(p/2+1)$ 和 $2\pi^{p/2}/\Gamma(p/2)$.]

1.49 设 $X_1, \cdots, X_n$ 为从总体 X 中取出的 i.i.d. 样本, $(X_{1:n}, \cdots, X_{n:n})$ 为其次序统计量, $P(0 \leqslant X \leqslant 1) = 1$. 若对任一 n, $X_{1:n}, X_{2:n} - X_{1:n}, \cdots, X_{n:n} - X_{n-1:n}$ 和 $1 - X_{n:n}$ 同分布, 则 $X \sim R(0,1)$.

1.50 设函数 $f : \mathbf{R} \mapsto \mathbf{R}$ 具有下述性质: 若 $X \sim N(\mu, \sigma^2)$, 则 $\mathrm{E}[f(X)] = g(\mu)$ 仅与 μ 有关, 与 σ 无关, 则 $f(x)$ 为 x 的线性函数.

第二章

点估计

本章讨论的主题是用 Wald 的统计决策理论的观点来研究参数的**点估计**. 设随机元 X 的样本空间为 $(\mathcal{X},\mathcal{B})$, 分布族为 $\{P_\theta,\theta\in\Theta\}$, 参数空间 Θ 为欧氏的. 要估计的对象一般是一个**参数函数** $g(\theta)$, 其中 $g(\cdot)$ **为在参数空间上定义的已知的向量值 Borel 可测函数**. 此时行动空间 $(\mathcal{A},\mathcal{B}_{\mathcal{A}})$ 为欧氏的, 其维数等于向量 $g(\theta)$ 的维数 m, 一般要求 $\mathcal{A}\supset g(\Theta)$. 按照适当选定的决策函数 δ, 一旦有了样本 X, 即可从 $\mathcal{A}$ 中挑出一个点去估计 $g(\theta)$. 这就是**点估计**名称的由来. 此时, 用于点估计的决策函数一般称为**估计**或**估计量**. 在非随机化情形, 它就是一个取值于 $(\mathcal{A},\mathcal{B}_{\mathcal{A}})$ 的统计量 δ, 当样本为 $X=x$ 时, 就用 $\delta(x)$ 估计 $g(\theta)$. 在随机化情形, 它是一个条件概率分布 $\delta(x,A)$, $(x,A)\in\mathcal{X}\times\mathcal{B}_{\mathcal{A}}$, 满足 1.3.3 节定义 1.7 中的两个条件. 一旦有了样本 x, 即以概率分布 $\delta(x,\cdot)$ 从 $\mathcal{A}$ 中选取一个点作为 $g(\theta)$ 的估计.

设在 $\Theta\times\mathbf{R}^m$ 上给定了损失函数 $L(\theta,a)$, 它可视为当参数真值为 θ, 而用 a 估计 $g(\theta)$ 时所受的损失. 假定 L 非负, 且对固定的 θ, 函数 $L(\theta,\cdot)$ 为 Borel 可测, 此时在非随机化情形, 估计量 δ 的风险函数为

$$R(\theta,\delta)=\mathrm{E}_\theta[L(\theta,\delta(X))]=\int_{\mathcal{X}}L\big(\theta,\delta(x)\big)dP_\theta(x),\ \theta\in\Theta. \tag{2.1}$$

当 $m=1$ 时, 若采用平方损失 $L(\theta,a)=(a-g(\theta))^2$, 则 $g(\theta)$ 的估计 $\delta(X)$ 的风险函数为 $R(\theta,\delta)=\mathrm{E}_\theta(\delta(X)-g(\theta))^2$, 称之为**估计 δ 的均方误差 (MSE)**. 若采用绝对值损失 $L(\theta,a)=|a-g(\theta)|$, 则 δ 的风险函数为 $R(\theta,\delta)=\mathrm{E}_\theta|\delta(X)-g(\theta)|$, 称之为估计 δ 的平均绝对误差.

在随机化情形, 风险函数有一个比较复杂的表达式

$$R(\theta,\delta)=\mathrm{E}_\theta\Big[\int_{\mathcal{A}}L(\theta,a)\delta(X,da)\Big]=\int_{\mathcal{X}}\Big[\int_{\mathcal{A}}L(\theta,a)\delta(x,da)\Big]dP_\theta(x),\ \theta\in\Theta \tag{2.2}$$

[参看 (1.25)].

按照 Wald 的观点, 风险愈小, 估计量愈好. 如果能找到某个估计量, 与其他估计量相比, 其风险一致地 (即对所有 $\theta \in \Theta$ 同时) 达到最小, 则它就是最佳的估计量. 但这样的估计量一般不存在, 因而只好放低标准. 通常有如下两种做法.

(1) 通过对估计量提出某种要求, 来限制可以采用的估计量的类, 希望在缩小了的估计类中, 能找出一个估计量, 其风险一致地达到最小. 提出的要求既要切合实际, 又要有严格的数学表达. 本章中涉及的限制原则有两个: 一是**无偏性**, 二是**同变性**.

(2) 放弃风险一致最小的准则, 改用某种综合性指标, 找出使该指标达到最小的估计量. 这种做法因其只涉及单一的数量指标, 解的存在性通常不成问题. 本章将要讨论的 **Bayes 估计**和 **minimax 估计**, 从某种意义上说, 就与第二种考虑有关.

2.1 无偏估计

简单地说, 一个估计量称为无偏的, 如果它的理论均值等于被估计的值. 严格地说, 我们有下面的定义.

定义 2.1 称 $\delta(X)$ 为 $g(\theta)$ 的一个**无偏估计**, 如果

$$\mathrm{E}_\theta\big[\delta(X)\big] = g(\theta), \text{ 对 } \forall\ \theta \in \Theta.$$

称 $g(\theta)$ 的一个随机化估计 $\delta(X, \cdot)$ 为无偏的, 如果

$$\mathrm{E}_\theta\left[\int_{\mathcal{A}} a\delta(X, da)\right] = g(\theta), \text{ 对 } \forall\ \theta \in \Theta.$$

无偏性是 Gauss 于 1821 年在他关于最小二乘理论的工作中引进的. 在各种准则中, 无偏性可能是点估计中用得最多的. 之所以如此, 一是在早期的统计著作中, 常常把无偏性放在显著的地位, 这是历史的原因. 此外, 它在数学上容易把握和处理也是一个重要原因. 而从应用的角度看, 当一个估计量需要在同一模型下多次重复使用时, 如果各次误差的后果可以正负相消, 则这种平均的效果是很有吸引力的. 即便不是多次反复使用, 一个没有系统误差的估计也许能让人更放心一些.

2.1.1 风险一致最小的无偏估计

首先我们给出如下定义.

定义 2.2 假定 $\hat{g}$ 是 $g(\theta)$ 的一个无偏估计, 如果对于 $g(\theta)$ 的任一无偏估计 δ, 总有

$$\mathrm{Var}_\theta(\hat{g}) \leqslant \mathrm{Var}_\theta(\delta), \quad \text{对 } \forall\ \theta \in \Theta,$$

则称 $\hat{g}$ 为 $g(\theta)$ 的方差一致最小的无偏估计 (UMVU 估计), 简称**最小方差无偏估计**.

无偏估计的方差就是它在平方损失之下的风险. 因此定义 2.2 可以推广到以下一般情形.

定义 2.3 假定 $\hat{g}$ 是 $g(\theta)$ 的一个无偏估计, 如果对于 $g(\theta)$ 的任一无偏估计 δ, 总有

$$R(\theta,\hat{g}) \leqslant R(\theta,\delta), \quad \text{对 } \forall\ \theta\in\Theta,$$

则称 $\hat{g}$ 为 $g(\theta)$ 的风险一致最小的无偏估计 (UMRU 估计), 简称**最小风险无偏估计**.

需要指出的是, 我们既可以举出无偏估计不存在的例子, 也可以举出存在无偏估计但不存在最小方差无偏估计的例子.

在点估计的风险理论中, 人们经常使用的是所谓的凸损失. 一个损失函数称为凸的, 如果对每个固定的 $\theta\in\Theta$, $L(\theta,\cdot)$ 为 $\mathcal{A}$ 上的一个凸函数, 此时 $\mathcal{A}$ 为 $\mathbf{R}^m$ 或是一个 m-维的开凸集. 形如 $L(\theta,a)=||g(\theta)-a||^2$ 的平方损失是凸损失中最重要的例子. 下面我们要在凸损失的假定之下, 在一般情况下寻求 $g(\theta)$ 的 MRU 估计.

今设 L 为凸损失, $\delta(X,da)$ 为 $g(\theta)$ 的任一随机化无偏估计. 据此可构作一非随机化估计 $\delta_0(X)=\int_{\mathcal{A}} a\delta(X,da)$. 其无偏性可由 δ 的无偏性得出, 且由定理 1.30,

$$R(\theta,\delta_0) \leqslant R(\theta,\delta), \quad \theta\in\Theta.$$

这表明如下结论.

定理 2.1 *在凸损失之下, 非随机化无偏估计的子类在无偏估计的类中是一个本质完全类.*

因此, 为寻求最小风险无偏估计, 不必考虑随机化估计.

由定理 1.31 立得如下定理.

定理 2.2 (无偏估计的 Rao-Blackwell 定理) *设 L 为凸损失, T 为充分统计量, $\delta(X)$ 为 $g(\theta)$ 的任一无偏估计, 则 $\hat{g}(T):=\mathrm{E}[\delta(X)|T]$ 也是 $g(\theta)$ 的无偏估计, 且对任一 $\theta\in\Theta$,*

$$R(\theta,\hat{g}(T)) \leqslant R(\theta,\delta). \tag{2.3}$$

若 $L(\theta,a)$ 对 a 为严格凸, $R(\theta,\delta)<\infty$, 则上面的严格不等号成立, 除非 $\delta(X)=\hat{g}[T(X)]$, a.s. P_θ.

引理 2.3 *仅依赖于完全统计量的无偏估计如果存在, 必 a.s. 唯一.*

证明 今设 T 为一完全统计量, 若 $\hat{g}_1(T)$ 和 $\hat{g}_2(T)$ 为 $g(\theta)$ 的两个无偏估计, 则由无偏性的定义, $\mathrm{E}_\theta[\hat{g}_1(T)-\hat{g}_2(T)]=0$ 对 $\forall\ \theta$, 因而由 T 的完全性, 在任何 P_θ 之下, $\hat{g}_1(T)=\hat{g}_2(T)$ a.s. 成立. ■

定理 2.4 (Lehmann-Scheffé 定理) *设样本空间为 $(\mathcal{X},\mathcal{B})$, $X\sim P_\theta,\theta\in\Theta$, T 为一完全充分统计量, 为估计 $g(\theta)$, 取损失函数 $L(\theta,\cdot)$ 为凸函数, 则*

(1) 设 $g^*(T)$ 为 $g(\theta)$ 的无偏估计, 则必为 $g(\theta)$ 的风险一致最小的无偏估计;

(2) 设 $\delta(X)$ 为 $g(\theta)$ 的任一无偏估计, 则 $\hat{g}(T):=\mathrm{E}[\delta(X)|T]$ 为 $g(\theta)$ 的风险一致最小的无偏估计;

(3) 设在严格凸的损失函数 $L(\theta,\cdot)$ 之下, $\delta(X)$ 为 $g(\theta)$ 的风险一致最小的无偏估计, 其风险函数处处有限, 则该估计必为 $T=T(X)$ 的函数, 因而也是唯一的.

证明 (1—2) 设 $\delta(X)$ 为 $g(\theta)$ 的任一无偏估计, 令 $\hat{g}(T)=\mathrm{E}[\delta(X)|T]$, 则它也是 $g(\theta)$ 的无偏估计. 由引理 2.3, $g^*(T)=\hat{g}(T)$, a.s. $\forall\ P_\theta$. 再由定理 2.2, $R(\theta,g^*(T))=R(\theta,\hat{g}(T))\leqslant R(\theta,\delta),\forall\ \theta,\forall\ \delta$.

(3) 设 $\delta(X)$ 为 $g(\theta)$ 的风险一致最小的无偏估计, 则由 Rao-Blackwell 定理, 刚刚定义的 $\hat{g}(T)=\mathrm{E}[\delta(X)|T]$ 也是 $g(\theta)$ 的风险一致最小的无偏估计, 两者具有有限且恒等的风险函数, 由 L 的严格凸性质, 知 $\delta(X)=\hat{g}[T(X)]$, a.s. $\forall\ P_\theta$. 再由引理 2.3, 即得其唯一性. ■

应当指出, 由 Lehmann-Scheffé 定理, 在具有完全充分统计量的样本分布族中, $g(\theta)$ 在任一严格凸损失下的最小风险无偏 (MRU) 估计与它的最小方差无偏 (UMVU) 估计完全一样.

以下通过若干例子, 说明如何根据 Lehmann-Scheffé 定理, 在一些分布族中寻求凸损失下 $g(\theta)$ 的 MRU 估计. 前提一是存在这样的估计, 二是能找到一个完全充分统计量 T.

有时直接就能看出某个 $\hat{g}(T)$ 为 $g(\theta)$ 的无偏估计, 则由上述定理 2.4, 这个 $\hat{g}(T)$ 即为 $g(\theta)$ 的 MRU 估计. 有时需要作出一些修正. 但在多数情形, 事情没有那么简单, 这时可求助于以下两种方法. 当然还可能有其他方法.

方法1. 假定损失是严格凸的, 则基于 T 的 MRU 估计 $\delta(T)$ 唯一地由以下的方程组确定:

$$\mathrm{E}_\theta[\delta(T)]=g(\theta),\ \theta\in\Theta. \tag{2.4}$$

方法2. 设 $\delta(X)$ 为 $g(\theta)$ 的一个无偏估计, 则由上述定理, $g(\theta)$ 在凸损失之下的 MRU 估计可通过对 $\delta(X)$ 在给定 T 的条件下取条件期望而得到.

在下面的例子中采用的都是凸损失.

例 2.1 先考虑一些简单的例子.

设 $X_1,\cdots,X_n$ 为取自两点分布 $B(1,p)$ 的 i.i.d. 样本, $n\geqslant 2, 0<p<1, q=1-p$. 为估计 X_1 的均值 p 和方差 pq, 记 $T=\sum_{i=1}^n X_i$, $S^2=\sum_{i=1}^n(X_i-\bar{X})^2/(n-1)$, 则 T 为完全充分统计量, $T\sim B(n,p)$, $S^2=T(n-T)/[n(n-1)]$ 为 T 的函数, 且 $T/n=\bar{X}$ 和 S^2 分别为 p 和 pq 的无偏估计, 故必为 MRU 估计.

在上述二项抽样的模型中, 不存在 $1/p$ 的无偏估计. 如其不然, 则由定理 2.2, 存在形如 $\delta(T)$ 的无偏估计. 利用 (2.4), 读者不难用反证法加以证明. 易见, p 接近于 0 时 p 的微小改变会引起 $1/p$ 的很大改变, 故要得到 $1/p$ 的可靠估计, p 越小观测数目应当越多. 负二项抽样很自然地符合这一要求. 这一抽样是在成功概率为 p 的重复的 Bernoulli 试验中, 试验一直进行到第 r 次成功为止. 用 Z 表示需要的试验次数. $Z\sim NB(r,p), 0<p<1$ (参看 1.2.3 节). 不难证明, Z 为完全充分统计量, $\mathrm{E}_p(Z)=r/p$, 故 Z/r 为 $1/p$ 的 MRU 估计.

今设 $X\sim P(\lambda),\lambda>0$, $X_1,\cdots,X_n$ 为取自上述 Poisson 总体的 i.i.d. 样本.

$T=\sum_{i=1}^n X_i\sim P(n\lambda)$ 且为完全充分统计量. 故 T/n 为 λ 的 MRU 估计.

下面对给定的非负整数 k, 寻求概率 $p_k(\lambda)=P_\lambda(X_1=k)=\mathrm{e}^{-\lambda}\lambda^k/k!$ 的 MRU 估计. 显然, $I(X_1=k)$ 是 $p_k(\lambda)$ 的一个无偏估计. 故所求的 MRU 估计为 $\hat{p}_k(T)=\mathrm{E}[I(X_1=k)|T]=P(X_1=k|T)$. 因为 $X_1\sim P(\lambda), \sum_{i=2}^n X_i\sim P\big((n-1)\lambda\big)$, 且两者独立, 故当 T 给定时, X_1 的条件分布为 $B(T,1/n)$. 因而 $p_k(\lambda)$ 的 MRU 估计为 $\hat{p}_k(T)=\binom{T}{k}(1/n)^k(1-1/n)^{T-k}$. 注意按照约定, 当 $T<k$ 时 $\binom{T}{k}=0$.

考虑两个均匀分布族, 一个是 $R(0,\theta),\theta>0$; 另一个是 $R(\theta_1,\theta_2),-\infty<\theta_1<\theta_2<\infty$. 设 $X_1,\cdots,X_n$ 分别为从这两个总体中抽取的 i.i.d. 样本, 则在凸损失之下, θ,θ_1,θ_2 的最小风险无偏估计可分别取为

$$\hat{\theta}=\frac{n+1}{n}X_{(n)},\quad \hat{\theta}_1=\frac{1}{n-1}\big(nX_{(1)}-X_{(n)}\big),\quad \hat{\theta}_2=\frac{1}{n-1}\big(nX_{(n)}-X_{(1)}\big),$$

其中, $X_{(1)}=\min_{1\leqslant i\leqslant n}X_i,\ X_{(n)}=\max_{1\leqslant i\leqslant n}X_i$.

例 2.2 设一有限总体 X 取值于集合 $A_\theta=\{1,2,\cdots,\theta\}$, θ 为一未知的正整数, $P_\theta(X=x)=1/\theta, x\in A_\theta$. 从中作无放回随机抽样, 得到样本 $X_1,\cdots,X_n, n<\theta$. 要在任一凸损失之下寻求 θ 的风险一致最小的无偏估计.

记 $x_{(n)}=\max_{1\leqslant i\leqslant n}x_i$. 由无放回抽样, 当 $x_1,\cdots,x_n$ 为不同的自然数时,

$$P_\theta(X_1=x_1,\cdots,X_n=x_n)=I\{x_{(n)}\leqslant\theta\}\Big/\binom{\theta}{n},$$

故由因子分解定理, $X_{(n)}:=\max_{1\leqslant i\leqslant n}X_i$ 为一充分统计量. 有

$$P_\theta(X_{(n)}=x)=\left\{\binom{x}{n}-\binom{x-1}{n}\right\}\Big/\binom{\theta}{n}=\binom{x-1}{n-1}\Big/\binom{\theta}{n},\ n\leqslant x\leqslant\theta.$$

此处约定当正整数 $x<n$ 时 $\binom{x}{n}=0$. 又若 $\mathrm{E}_\theta[f(X_{(n)})]=0,\forall\ \theta\geqslant n$, 则容易得出 $P_\theta\{f(X_{(n)})\neq 0\}=0$, 故 $X_{(n)}$ 也是完全统计量. 由于

$$\begin{aligned}\mathrm{E}_\theta(X_{(n)})\ &=\sum_{x=n}^{\theta}x\binom{x-1}{n-1}\Big/\binom{\theta}{n}=n\sum_{x=n}^{\theta}\binom{x}{n}\Big/\binom{\theta}{n}\\&=n\binom{\theta+1}{n+1}\Big/\binom{\theta}{n}=\frac{n}{n+1}(\theta+1),\end{aligned}$$

故据 Lehmann-Scheffé 定理,

$$\hat{\theta}:=\frac{n+1}{n}X_{(n)}-1$$

即为 θ 的风险一致最小的无偏估计.

该例取自陈家鼎, 孙山泽, 李东风, 刘力平 (2006) 第 23—27 页. 书中讲到第二次世界大战期间, 盟军曾用掌握的德军的某种坦克的 “序列号” 估计该种坦克的总数, 效果很好, 书中介绍了当年采用的估计. 另外, 郑忠国, 童行伟, 赵慧 (2012), 找到一个更好的估计, 后者是无偏的而且方差最小. 这一估计正是上面的 $\hat{\theta}$.

例 2.3 (单边截断族) 设 $X\sim f(x,\theta)=h(x)I(a<x<\theta)\,dx/H(\theta), a<\theta<b$,

其中 h 为区间 (a,b) 上给定的恒正的 Borel 可测函数, $0<H(\theta)=\int_a^\theta h(t)dt<\infty$, $X_1,\cdots,X_n$ 为从上述总体中抽取的 i.i.d. 样本. 设 g 为定义于 (a,b) 上的已知函数. 直接利用 (2.4), 我们将证明, 在任一严格凸损失之下, $g(\theta)$ 的 MRU 估计存在的充要条件为:

(1) 对任何 $c\in(a,b)$, $d[g(t)H^n(t)]/dt$ 在区间 (a,c) 上 Lebesgue 可积;

(2) $\lim_{t\downarrow a}[H^n(t)g(t)]=0$.

为此, 令 $T=X_{(n)}=\max_{1\leqslant i\leqslant n}X_i$, 则 T 为完全充分统计量, 其密度为

$$f_n(t,\theta)=nH^{n-1}(t)h(t)/H^n(\theta),\quad a<t<\theta.$$

若 $\delta(T)$ 为 $g(\theta)$ 的无偏估计, 则由无偏性条件 (2.4), 有

$$\int_a^\theta nH^{n-1}(t)h(t)\delta(t)dt=g(\theta)H^n(\theta),\quad a<\theta<b.$$

因而条件 (1) 成立, 且

$$nH^{n-1}(t)h(t)\delta(t)=d[g(t)H^n(t)]/dt=g'(t)H^n(t)+nH^{n-1}(t)h(t)g(t)$$

对 $t\in(a,b)$ a.e. L 成立, 此处 L 为 Lebesgue 测度. 故

$$\delta(T)=g(T)+g'(T)H(T)/[nh(T)],\text{a.s.}.$$

此时由条件 (1),

$$\begin{aligned}\mathrm{E}_\theta[\delta(T)]&=H^{-n}(\theta)\int_a^\theta nH^{n-1}(t)h(t)\delta(t)dt\\&=H^{-n}(\theta)\int_a^\theta\frac{d}{dt}\big\{H^n(t)g(t)\big\}dt\\&=H^{-n}(\theta)\big\{H^n(\theta)g(\theta)-\lim_{t\downarrow a}[H^n(t)g(t)]\big\}.\end{aligned}$$

故要使 $\delta(T)$ 为 $g(\theta)$ 的无偏估计, 当且仅当条件 (2) 成立. 由 Lehmann-Scheffé 定理, 在此条件下, $\delta(T)=g(T)+g'(T)H(T)/[nh(T)]$ 即为 $g(\theta)$ 的 MRU 估计.

Tate (1959) 研究过单边截断族中 $g(\theta)$ 的最小方差无偏估计, 求出了这一估计的表达式 [也可参看 Zacks (1971)]. Tate 所用的方法是先寻求只依赖于 X_1 的无偏估计. 由于 $g(\theta)$ 的无偏估计存在的条件事实上依赖于样本量 n, 他所加的条件要比此处的强. 双边截断族的情形较单边情形复杂, 可参看卢昆亮, 赵林城 (1982).

例 2.4 设 $X_1,\cdots,X_n$ 为从正态总体 $N(a,\sigma^2),-\infty<a<\infty,\sigma>0$ 中抽取的 i.i.d. 样本. 用 Φ 和 φ 分别表示标准正态 $N(0,1)$ 的分布和密度函数. 要对给定的 u, 找出凸损失下 $p=P_\theta(X_1\leqslant u)=\Phi[(u-a)/\sigma]$ 的 MRU 估计, 其中 $\theta=(a,\sigma)$. 著名数学家 Kolmogorov 于 1950 年解决了这个问题.

先考虑 $\sigma^2=1$ 的简单情形. 此时 $\bar X=\sum_{i=1}^n X_i/n$ 是完全充分统计量, $I(X_1\leqslant u)$ 为 p 的一个无偏估计, 故由定理 2.4, $P(X_1\leqslant u|\bar X)$ 即为所求的 MRU 估计. 因为 $X_1-\bar X$ 的分布与参数无关, 由 Basu 定理, 它与 $\bar X$ 独立, 故由 $X_1-\bar X\sim N(0,(n-1)/n)$,

有

$$\begin{aligned} P(X_1 \leqslant u|\bar{X}=\bar{x}) &= P(X_1-\bar{X} \leqslant u-\bar{x}|\bar{X}=\bar{x}) \\ &= P(X_1-\bar{X} \leqslant u-\bar{x}) \\ &= \Phi[\sqrt{n/(n-1)}(u-\bar{x})], \end{aligned}$$

它就是 p 的 MRU 估计.

一个有关的问题是估计 X_1 在 u 点的概率密度 $p_a^{X_1}(u)=\varphi(u-a)$. 我们将证明, 它的 MRU 估计正是给定 $\bar{X}$ 时 X_1 在 u 点的条件密度 $\delta(\bar{X})=p^{X_1|\bar{X}}(u)$. 因为 $(X_1,\bar{X})$ 的联合密度为 $f_a(u,v)=p^{X_1|\bar{X}}(u)p_a^{\bar{X}}(v)$, 故

$$p_a^{X_1}(u)=\int_{-\infty}^{\infty} p^{X_1|\bar{X}}(u)p_a^{\bar{X}}(v)dv.$$

这表明 $\delta(\bar{X})=p^{X_1|\bar{X}}(u)$ 为 $p_a^{X_1}(u)$ 的一个无偏估计. 对刚刚得到的方程

$$P(X_1 \leqslant u|\bar{x})=\Phi[\sqrt{n/(n-1)}(u-\bar{x})]$$

取导数, 即得

$$p^{X_1|\bar{X}}(u)=\sqrt{n/(n-1)}\varphi[\sqrt{n/(n-1)}(u-\bar{x})].$$

这就是 $p_a^{X_1}(u)=\varphi(u-a)$ 的 MRU 估计.

现在假定 a 和 σ 都是未知的, 则 $(\bar{X},S)$ 为完全充分统计量. 跟前面完全一样, $p=P_\theta(X_1 \leqslant u)=\Phi[(u-a)/\sigma]$ 和密度 $p_\theta^{X_1}(u)=(1/\sigma)\varphi[(u-a)/\sigma]$ 的 MRU 估计分别为 $P(X_1 \leqslant u|\bar{X},S)$ 和给定 $\bar{X},S$时 X_1 在 u 点的条件密度, 此处, $S^2=\sum_{i=1}^n(X_i-\bar{X})^2$. 因为 $(X_1-\bar{X})/S$ 的分布与参数无关, 由 Basu 定理, 它与 $(\bar{X},S)$ 独立, 与前面一样, 有

$$\begin{aligned} &P(X_1 \leqslant u|\bar{x},s)=P\Big(\sqrt{\frac{n}{n-1}}\frac{X_1-\bar{X}}{S} \leqslant \sqrt{\frac{n}{n-1}}\frac{u-\bar{x}}{s}\Big) := P(V \leqslant v_0), \\ &p^{X_1|\bar{X},S}(u)=\sqrt{\frac{n}{n-1}}\frac{f(v_0)}{s}, \end{aligned}$$

其中 $V=\sqrt{n/(n-1)}(X_1-\bar{X})/S$, $v_0=\sqrt{n/(n-1)}(u-\bar{x})/s$, 而 f 为 V 的密度.

为求 V 的分布, 只需假定 $X_1 \sim N(0,1)$. 作由 $(X_1,\cdots,X_n)^\tau$ 到 $(Y_1,\cdots,Y_n)^\tau$ 的正交变换, 其中 $Y_1=\sqrt{n}\bar{X}$, $Y_2=\sqrt{n/(n-1)}(X_1-\bar{X})$. 由引理 1.16, $Y_1,\cdots,Y_n$ 为 i.i.d. 正态 $N(0,1)$ 样本, 且 $S^2=\sum_{i=2}^n Y_i^2$, $V=Y_2/S=Y_2/(\sum_{i=2}^n Y_i^2)^{1/2}$. 故由推论 1.17 (2), $V^2 \sim Be(1/2,n/2-1)$. 由此即可算得 V 的密度函数为

$$f(v)=\frac{\Gamma\big((n-1)/2\big)}{\sqrt{\pi}\Gamma(n/2-1)}(1-v^2)^{n/2-2}I(|v|<1),$$

故 p 与 $p_\theta^{X_1}(u)$ 的 MRU 估计分别为

$$P(V \leqslant v_0)=\int_{-1}^{v_0}\frac{\Gamma\big((n-1)/2\big)}{\sqrt{\pi}\Gamma(n/2-1)}(1-t^2)^{n/2-2}dtI(|v_0|<1)+I(v_0 \geqslant 1)$$

与

$$\sqrt{\frac{n}{n-1}}\frac{f(v_0)}{s}=\frac{\sqrt{n}\Gamma\big((n-1)/2\big)}{\sqrt{(n-1)\pi}\Gamma(n/2-1)}\frac{(1-v_0^2)^{n/2-2}}{s}I(|v_0|<1),$$

其中 $v_0=\sqrt{n/(n-1)}(u-\bar{x})/s$.

例 2.5 设 $X_1,\cdots,X_n$ 为从指数分布总体 $E(a,b),-\infty<a<\infty,b>0$ 中抽取的 i.i.d. 样本. 要对给定的 u, 找出凸损失下可靠性 $P_{a,b}(X_1>u)$ 的 MRU 估计.

用 $X_{1:n}\leqslant X_{2:n}\leqslant\cdots\leqslant X_{n:n}$ 记 $X_1,\cdots,X_n$ 的次序统计量. 由 1.6.2 节例 1.13, 可知 $(X_{1:n},T)$ 是完全充分统计量, 其中 $T=\sum_{i=1}^n(X_i-X_{1:n})=\sum_{i=2}^n(X_{i:n}-X_{1:n})$. 由定理 2.4, $P(X_1>u|X_{1:n},T)$ 即为所求的 MRU 估计. 上述条件概率与参数无关, 不妨设 $P=P_{0,1}$.

因为 $(X_1-X_{1:n})/T$ 的分布与参数无关, 由 Basu 定理, 它与 $(X_{1:n},T)$ 独立. 因此

$$\begin{aligned}P(X_1>u|x_{1:n},t)&=P\Big(\frac{X_1-X_{1:n}}{T}>\frac{u-x_{1:n}}{t}\Big|X_{1:n}=x_{1:n},T=t\Big)\\&=P\Big(\frac{X_1-X_{1:n}}{T}>\frac{u-x_{1:n}}{t}\Big):=P(V>\xi).\end{aligned}$$

此处, $V=(X_1-X_{1:n})/T$, $\xi=(u-x_{1:n})/t$. 因为 $0\leqslant V\leqslant 1$, 故当 $\xi<0$ 时, $P(V>\xi)=1$; 当 $\xi\geqslant 1$ 时, $P(V>\xi)=0$. 今设 $0\leqslant\xi<1$.

给定 $X_{1:n}$, $(X_{2:n}-X_{1:n},\cdots,X_{n:n}-X_{1:n})$ 与取自指数分布 $E(0,1)$ 的 i.i.d. 样本 $Y_1,\cdots,Y_{n-1}$ 的次序统计量 $(Y_{1:n-1},\cdots,Y_{n-1:n-1})$ 同分布. 由对称性,

$$\begin{aligned}P\Big(V>\xi\Big|X_{1:n}\Big)&=\frac{1}{n}\sum_{i=1}^n P\Big(\frac{X_i-X_{1:n}}{T}>\xi\Big|X_{1:n}\Big)\\&=\frac{1}{n}\sum_{i=1}^n P\Big(\frac{X_{i:n}-X_{1:n}}{T}>\xi\Big|X_{1:n}\Big)\\&=\frac{1}{n}\sum_{i=2}^n P\Big(\frac{X_{i:n}-X_{1:n}}{T}>\xi\Big|X_{1:n}\Big)\\&=\frac{1}{n}\sum_{i=1}^{n-1} P\Big(\frac{Y_{i:n-1}}{\sum_{k=1}^{n-1}Y_k}>\xi\Big)\\&=\frac{1}{n}\sum_{i=1}^{n-1} P\Big(\frac{Y_i}{\sum_{k=1}^{n-1}Y_k}>\xi\Big)\\&=\frac{n-1}{n}P\Big(\frac{Y_1}{\sum_{k=1}^{n-1}Y_k}>\xi\Big).\end{aligned}$$

但 $Z:=Y_1/\sum_{k=1}^{n-1}Y_k\sim Be(1,n-2)$, 故由以上的推导, 当 $0\leqslant\xi<1$ 时,

$$\begin{aligned}P(X_1>u|x_{1:n},t)=P(V>\xi)&=\frac{n-1}{n}P(Z>\xi)\\&=\frac{n-1}{n}\frac{1}{n-2}\int_\xi^1(1-t)^{n-3}dt\\&=\frac{n-1}{n}(1-\xi)^{n-2}.\end{aligned}$$

综上所述, 可靠性 $P_{a,b}(X_1 > u)$ 的 MRU 估计为

$$\begin{aligned} &P(X_1 > u|X_{1:n}, T) \\ &= \frac{n-1}{n}\Big(1 - \frac{u - X_{1:n}}{T}\Big)^{n-2} I\Big(0 \leqslant \frac{u - X_{1:n}}{T} \ < \ 1\Big) + I(X_{1:n} > u). \end{aligned}$$

在本节结束之前, 我们还要介绍关于 UMVU 估计的一个定理. 它和定理 2.5 都是在 Lehmann 和 Scheffé (1950, 1955, 1956) 中提出的. 为了叙述的方便, 我们称统计量 U 为 0 的一个无偏估计, 如果 $\mathrm{E}_\theta(U) = 0$, 对 $\forall\ \theta \in \Theta$.

定理 2.5 *设 δ 为 $g(\theta)$ 的无偏估计, 对 $\forall\ \theta \in \Theta$, $\mathrm{Var}_\theta(\delta) < \infty$, 则 δ 为 $g(\theta)$ 的 UMVU 估计的充要条件为: 对 0 的任一无偏估计 U, 如果它在 P_{θ^*} 之下的方差有限, 则必有 $\mathrm{E}_{\theta^*}(\delta U) = 0$.*

(注意, $\mathrm{E}_{\theta^*}(\delta U) = \mathrm{Cov}_{\theta^*}(\delta, U)$.)

证明 **必要性.** 假定 δ 为 $g(\theta)$ 的 UMVU 估计, 则对 0 的任一无偏估计 U 和任意的实数 λ, $\delta + \lambda U$ 仍为 $g(\theta)$ 的一个无偏估计, 因而

$$\mathrm{Var}_\theta(\delta + \lambda U) \geqslant \mathrm{Var}_\theta(\delta), \quad \text{对所有 } \lambda,$$

即

$$\lambda^2 \mathrm{Var}_\theta U + 2\lambda \mathrm{E}_\theta(\delta U) \geqslant 0, \quad \text{对所有 } \lambda.$$

今设 $\mathrm{Var}_\theta(U) < \infty$. 如果 $\mathrm{E}_\theta(\delta U) \neq 0$, 则由 Schwarz 不等式, $0 < |\mathrm{E}_\theta(\delta U)|^2 = |\mathrm{Cov}_\theta(\delta, U)|^2 \leqslant \mathrm{Var}_\theta(\delta)\mathrm{Var}_\theta(U)$, 故 $\mathrm{Var}_\theta(U) > 0$. 此时上式左方的这个 λ 的二次式有两个不同的实根 $\lambda = 0$ 和 $\lambda = -2\mathrm{E}_\theta(\delta U)/\mathrm{Var}_\theta U$, 因而必可对某些 λ 取到负值, 这个矛盾就证明了命题的必要性.

充分性. 今设 δ 为 $g(\theta)$ 的方差有限的无偏估计, 且满足与 0 的任一无偏估计有关的条件. 为证它为 UMVU 估计, 设 δ_1 为 $g(\theta)$ 的任一无偏估计. 任取 $\theta \in \Theta$. 如果 $\mathrm{Var}_\theta(\delta_1) = \infty$, 则结论自动成立. 今设 $\mathrm{Var}_\theta(\delta_1) < \infty$, 则 $\delta - \delta_1$ 为 0 的无偏估计, $\mathrm{Var}_\theta(\delta - \delta_1) < \infty$, 由定理条件, $\mathrm{E}_\theta[\delta(\delta - \delta_1)] = 0$, 因而 $\mathrm{Var}_\theta\delta = \mathrm{Cov}_\theta(\delta, \delta_1)$, 由 Schwarz 不等式即得 $\mathrm{Var}_\theta\delta \leqslant \mathrm{Var}_\theta\delta_1$. ■

定理 2.5 特别适用于样本分布族没有完全统计量的情形, 可用来甄别哪些参数函数有 UMVU 估计而哪些没有, 参看习题 2.2.

2.1.2 Cramér-Rao 不等式

2.1.1 节讨论的问题是, 在平方损失之下, 寻求 $g(\theta)$ 的方差最小的无偏估计. 一般不难设想, 在一定的样本量之下, 这个方差不可能任意接近于 0, 应该有一个与 θ 有关的大于 0 的下界. 本小节将研究与这个下界有关的问题, 其基本结果是 Rao 和 Cramér 在 1945—1946 年间分别建立的一个不等式. 后来人们称其为 **Cramér-Rao 不等式**, 或简称为 **C-R 不等式**.

1. 单参数情形的 C-R 不等式

设样本分布为 $f(x,\theta)d\mu(x)$, μ 为样本空间 $(\mathcal{X}, \mathcal{B})$ 上的一个 σ-有限测度. 参数

$\theta \in \Theta, \Theta$ 为 $\mathbf{R}^1$ 上的一个开区间. 引进以下的正则性条件:

(1) 对 $\forall\, x \in \mathcal{X}, \theta \in \Theta$, $f(x,\theta) > 0$.

(2) 对 $\forall\, \theta \in \Theta$, $\partial f(x,\theta)/\partial\theta$ 存在, 且 $\int_{\mathcal{X}} \{\partial f(x,\theta)/\partial\theta\} d\mu(x) = 0$.

(3) 对任一 $\theta \in \Theta$, 有

$$I(\theta) := \mathrm{E}_\theta[S(X,\theta)]^2 > 0,$$

其中 $S(x,\theta) = \partial \log f(x,\theta)/\partial\theta$.

定理 2.6 (C-R 不等式) *设条件 (1)—(3) 成立, 且设 $T(X)$ 为 $g(\theta)$ 的一个无偏估计, 满足条件*

(4) $\int_{\mathcal{X}} T(x) f(x,\theta) d\mu(x) \equiv g(\theta)$ *可在积分号下对 θ 求导.*

则对任一 $\theta \in \Theta$,

$$\mathrm{Var}_\theta[T(X)] \geqslant [g'(\theta)]^2 / \mathrm{E}_\theta[S(X,\theta)]^2. \tag{2.5}$$

条件 (1) 要求样本分布族有共同支撑, 即对任一 θ, $P_\theta(N) = 0$ 当且仅当 $\mu(N) = 0$. 这就排除了形如 $\{R(0,\theta),\ \theta > 0\}$ 的截断族. 因为对 $\forall\,\theta$ 总有 $\int_{\mathcal{X}} f(x,\theta) d\mu(x) = 1$, 条件 (2) 等价于要求可在此积分号下对 θ 求导. 条件 (4) 蕴含了 $g'(\theta)$ 的存在性.

证明 不妨设 $\mathrm{Var}_\theta[T(X)]$ 和 $\mathrm{E}_\theta[S(X,\theta)]^2$ 都有限, 否则 (2.5) 自然成立. 因为条件 (2) 蕴含 $\mathrm{E}_\theta[S(X,\theta)] = 0$, 故

$$\mathrm{Cov}_\theta[T(X), S(X,\theta)] = \mathrm{E}_\theta[T(X)S(X,\theta)] = \int_{\mathcal{X}} T(x) \frac{\partial f(x,\theta)}{\partial\theta} d\mu(x).$$

由条件 (4),

$$\mathrm{Cov}_\theta[T(X), S(X,\theta)] = \frac{\partial}{\partial\theta} \int_{\mathcal{X}} T(x) f(x,\theta) d\mu(x) = g'(\theta).$$

用 Schwarz 不等式, 有

$$\mathrm{Var}_\theta[T(X)]\ \mathrm{E}_\theta[S(X,\theta)]^2 \geqslant [g'(\theta)]^2, \tag{2.6}$$

注意到条件 (3) 即得 (2.5). ■

如果 $g(\theta)$ 的所有无偏估计都满足上述的条件 (4), 而其中某个 $\hat{g}$ 的方差达到了 C-R 下界, 则 $\hat{g}$ 就是 $g(\theta)$ 的 UMVU 估计. 但在许多场合, 即使某个估计是 UMVU 的, 它的方差也常常达不到 C-R 下界. 因此, 这一不等式在寻求 UMVU 估计中的作用是很有限的. 下面的定理也说明了这一点. 但这个不等式在统计理论方面尚有其他作用, 并不限于寻求 UMVU 估计. 为了下面的讨论, 我们还对分布族引进以下的假定条件:

(5) 对任一给定的 $x \in \mathcal{X}$, $\partial f(x,\theta)/\partial\theta$ 为 Θ 上的连续函数.

显然, 若 $f(x,\theta) = C(\theta)\exp\{\theta T(x)\}h(x)$ 为一单参数指数族, 则条件 (1)—(3) 成立, 且对 $g(\theta)$ 的任一无偏估计 T, 条件 (4) 也成立. 事实上, 在一定条件下, 指数族还是能使 C-R 不等式等号成立的唯一的一个分布族.

定理 2.7 *假定条件 (1)—(5) 成立, 其中 $T(X)$ 为 $g(\theta)$ 的一个无偏估计, $g(\theta)$ 不恒为常数, $\mathrm{Var}_\theta(T) < \infty, \theta \in \Theta$, 则 (2.5) 中的等号对所有 $\theta \in \Theta$ 成立的充要条件*

是: 存在 $N \in \mathcal{B}, \mu(N)=0$, 当 $x \notin N, \theta \in \Theta$ 时有

$$f(x,\theta)=C(\theta)h(x)\exp\{Q(\theta)T(x)\}, \tag{2.7}$$

此处 $C(\theta)>0$, Q 在 Θ 上可微, 严格单调, $h(x)$ 恒正.

证明 充分性易证, 留给读者作为练习. 今证必要性. 由 Schwarz 不等式, (2.5) 也即 (2.6) 中等号成立的充要条件为: 存在 $\beta(\theta)$, $\gamma(\theta)$ 不全为 0, 使

$$\beta(\theta)S(x,\theta)=\gamma(\theta)[T(x)-g(\theta)], \quad \text{a.e. } \mu,$$

此处用到了分布族有共同支撑的假定. 对 $\forall\ \theta$, $\beta(\theta),\gamma(\theta)$ 均不为 0, 否则会与 $g(\theta)$ 不恒为常数的假定或条件 (3) 矛盾. 因而存在恒不为 0 的函数 $\alpha(\theta)$, 及一个 μ-零集 K_θ, 使得当 $x \notin K_\theta, \theta \in \Theta$ 时, 总有

$$S(x,\theta)=\alpha(\theta)[T(x)-g(\theta)]. \tag{2.8}$$

在参数空间 Θ 上引进 Borel 域 $\mathcal{C}$, 并用 L 表示 $(\Theta,\mathcal{C})$ 上的 Lebesgue 测度. 进一步可引进乘积空间和乘积测度 $\mu \times L$. 由 Fubini 定理, 存在一个 μ-零集 N, 当 $x \notin N$ 时, (2.8) 按 L 测度对 a. e. 的 θ 成立.

取 $x_1 \notin N, x_2 \notin N$, 使 $T(x_1) \neq T(x_2)$, 则存在一个 L-零测集 N^x (可与 x_1, x_2 有关), 使得当 $x \notin N$ 且 $\theta \notin N^x$ 时, (2.8) 在 x, x_1, x_2 处都成立. 定义

$$a(\theta)=\frac{S(x_2,\theta)-S(x_1,\theta)}{T(x_2)-T(x_1)}, \quad \theta \in \Theta.$$

由条件 (5), 上式定义的 a 为 Θ 上的连续函数. 在 $\Theta - N^x$ 上函数 a 与 α 取相同的值, 故当 $x \notin N$ 且 $\theta \in \Theta - N^x$ 时,

$$S(x,\theta)=a(\theta)[T(x)-g(\theta)]. \tag{2.9}$$

因为 $\Theta - N^x$ 在 Θ 中稠密, 且上式左右均为 Θ 上的连续函数, 因而上式当 $x \in \mathcal{X}-N, \theta \in \Theta$ 时成立! 取定 $\theta_0 \in \Theta$, 记

$$f(x,\theta_0)=h(x),\ Q(\theta)=\int_{\theta_0}^{\theta} a(t)dt, C(\theta)=\exp\left\{-\int_{\theta_0}^{\theta} a(t)g(t)dt\right\},$$

由 (2.9) 即得 (2.7). 又因为函数 a 连续且恒不为 0, 在 Θ 上必保持同号, 故函数 Q 可微且严格单调. ■

此外, 我们还有如下推论.

推论 2.8 在满足定理条件的指数族中, 当且仅当

$$g(\theta)=a\mathrm{E}_\theta[T(X)]+b, \quad a,b \text{ 为常数}$$

时, 才有处处达到 C-R 下界的无偏估计 $\hat{g}(X)=aT(X)+b$.

我们把这一推论的证明留给读者.

一维正态总体 $N(\theta,1)$, 指数分布 $E(0,\theta)$, 二项分布 $B(n,\theta)$, Poisson 分布 $P(\theta)$ 中的未知参数 θ 是能用 C-R 方法求 UMVU 估计的几个代表性的例子, 它们都满足定理 2.6 的条件. 对分布 $E(0,\theta)$, 按定理 2.4, $g(\theta)=\theta^2$ 的基于样本 X 的 UMVU 估计即为 $X^2/2$, 但它的方差达不到 C-R 下界, 其原因已在推论 2.8 中指出.

有些文献研究了 C-R 不等式的进一步精确化, 本书就不作介绍了.

2. Fisher 信息函数

设 $\mathcal{P}=\{P_\theta,\theta\in\Theta\}$ 为 X 的分布族. 在有关的正则条件下, 由 C-R 不等式, θ 的任一无偏估计的方差都不能小于 $I^{-1}(\theta)$, 其中

$$I(\theta)=\mathrm{E}_\theta[S(X,\theta)]^2=\mathrm{E}_\theta[\partial\log f(X,\theta)/\partial\theta]^2,\ \theta\in\Theta. \tag{2.10}$$

以此为由头可以引进 $I(\theta)$ 这个量. $I(\theta)$ 称为分布族 $\mathcal{P}$ 的 **Fisher 信息函数**. 它是 Fisher 提出的, 可用作有关分布族所含信息量的一种度量. 有时将它更明确地记为 $I^X(\theta)$. $I(\theta)$ 越大, 可能达到的方差下界就越小, 就越容易对 θ 作出精度高的估计, 即分布族 $\mathcal{P}$ 所含的 θ 的 "信息量" 越多. $I(\theta)$ 不仅与分布族有关, 而且随着 θ 的不同而不同, 反映在不同参数之下估计的难易程度. 考虑到 C-R 的这一下界 $I^{-1}(\theta)$ 一般是达不到的, 上述解释看来仍有欠缺. 以后我们将会看到, $I(\theta)$ 在点估计的渐近理论中有重要作用, 仅此一点, 就应该引起我们足够的重视.

下面我们不加证明地列举 $I(\theta)$ 的几个性质.

(1) $I(\theta)$ 是分布族本身的属性, 与测度 μ 的选取无关.

(2) 设 X,Y 独立, 它们的分布族皆满足本小节开头的正则性条件 (1) 和 (2), 则 $I^{(X,Y)}(\theta)=I^X(\theta)+I^Y(\theta)$. 因此, 若样本 $X=(X_1,\cdots,X_n)$, 其中 $X_1,\cdots,X_n$ i.i.d., 则 $I^X(\theta)=nI^{X_1}(\theta)$.

关于这两点的证明, 读者不难自己写出.

另外, 设 X 为一样本, 而 T 为一统计量, 则在很一般的条件下, 有 $I^T(\cdot)\leqslant I^X(\cdot)$, 且在一定的条件下可以证明, $I^T(\cdot)\equiv I^X(\cdot)$ 成立的充要条件是: T 为一充分统计量. 关于这一命题的确切陈述及其证明, 可参看陈希孺 (1981, 1997); 成平, 陈希孺, 陈桂景, 吴传义 (1985).

2.1.3 多个参数的情况

设参数 $\theta=(\theta_1,\cdots,\theta_k)^\tau$ 为 k-维的, $k\geqslant 2$, 要估计它的一个 r-维向量函数 $g(\theta)=\big(g_1(\theta),\cdots,g_r(\theta)\big)^\tau$, $r\leqslant k$. 与一维情形相同, 估计 $\hat{g}=\big(\hat{g}_1,\cdots,\hat{g}_r\big)^\tau$ 称为**无偏的**, 若对 $\forall\,\theta$, $\mathrm{E}_\theta(\hat{g})=g(\theta)$, 或 $\mathrm{E}_\theta(\hat{g}_j)=g_j(\theta)$, $1\leqslant j\leqslant r$. 称 $\hat{g}$ 为 $g(\theta)$ 的 **UMVU 估计**, 若它的每个分量 $\hat{g}_j$ 是 $g_j(\theta)$ 的 UMVU 估计. 我们有下面的引理.

引理 2.9 设 $\hat{g}_j$ 是 $g_j(\theta)$ 的 UMVU 估计, 且对任何 $\theta\in\Theta$, $\mathrm{Var}_\theta(\hat{g}_j)<\infty, 1\leqslant j\leqslant r$, 则对任何常数 $c_1,\cdots,c_r$, $\sum_{j=1}^r c_j\hat{g}_j$ 为 $\sum_{j=1}^r c_jg_j(\theta)$ 的 UMVU 估计.

证明 由假定, $g^*:=\sum_{j=1}^r c_j\hat{g}_j$ 为 $\sum_{j=1}^r c_jg_j(\theta)$ 的无偏估计, 方差有限. 设 U 为 0 的无偏估计, 则因为 $\hat{g}_j$ 是 $g_j(\theta)$ 的 UMVU 估计, 由定理 2.5, $\mathrm{Var}_\theta(U)<\infty$ 蕴含 $\mathrm{E}_\theta(\hat{g}_jU)=0$, $1\leqslant j\leqslant r$, 故 $\mathrm{E}_\theta(g^*U)=0$. 由定理 2.5, g^* 即为 $\sum_{j=1}^r c_jg_j(\theta)$ 的 UMVU 估计. ■

下述定理 2.10 表明, 多维参数函数的 UMVU 估计在以协方差阵为比较准则时仍然具有最优性. 此处对两个同阶半正定方阵 A,B, 称 $A\geqslant B$, 若 $A-B\geqslant 0$, 即 $A-B$ 为半正定. 称 B 在某个半正定方阵的类中是最小的, 如果对该类中的每个方

阵 A, 都有 $A \geqslant B$.

定理 2.10 设 $g(\theta) = \big(g_1(\theta), \cdots, g_r(\theta)\big)^\tau$, $\hat{g}_j$ 是 $g_j(\theta)$ 的 UMVU 估计, 且对任一 $\theta \in \Theta$, $\mathrm{Var}_\theta(\hat{g}_j) < \infty, 1 \leqslant j \leqslant r$. 又设 $\delta = (\delta_1, \cdots, \delta_r)^\tau$ 为 $g(\theta)$ 的任一无偏估计, 对任何 $\theta \in \Theta$, $\mathrm{Var}_\theta(\delta_j) < \infty, 1 \leqslant j \leqslant r$, 则有

$$\mathrm{Cov}_\theta(\hat{g}) \leqslant \mathrm{Cov}_\theta(\delta), \ \theta \in \Theta. \tag{2.11}$$

证明 设 $\hat{g}_j$ 是 $g_j(\theta)$ 的 UMVU 估计, 方差有限, $1 \leqslant j \leqslant r$. 按引理 2.9, 对任何 $c = (c_1, \cdots, c_r)^\tau$, $\sum_{j=1}^r c_j \hat{g}_j$ 为 $\sum_{j=1}^r c_j g_j(\theta)$ 的 UMVU 估计. 今若 $\delta = (\delta_1, \cdots, \delta_r)^\tau$ 为 $g(\theta)$ 的无偏估计, 则 $\sum_{j=1}^r c_j \delta_j$ 为 $\sum_{j=1}^r c_j g_j(\theta)$ 的无偏估计, 故应有

$$\mathrm{Var}_\theta\Big(\sum_{j=1}^r c_j \hat{g}_j\Big) \leqslant \mathrm{Var}_\theta\Big(\sum_{j=1}^r c_j \delta_j\Big), \quad \forall\, \theta \in \Theta.$$

由于 $\hat{g}_1, \cdots, \hat{g}_r$ 和 $\delta_1, \cdots, \delta_r$ 的方差都有限, 上式可写为

$$c^\tau \mathrm{Cov}_\theta(\hat{g}) c \leqslant c^\tau \mathrm{Cov}_\theta(\delta) c.$$

由于此式对任何 r-维向量 c 都成立, 按定义即得 (2.11). ■

由定理 2.10, 结合定理 2.4, 我们有如下推论.

推论 2.11 设 $g(\theta)$ 为 r-维向量函数, T 为一完全充分统计量, $\hat{g}$ 为基于 T 的 $g(\theta)$ 的无偏估计, 且 $\hat{g}$ 的各分量都有有限的方差. 定义如下的估计类

$$\mathcal{E} = \{\delta:\ \delta\text{为 } g(\theta) \text{ 的无偏估计, 其各分量方差有限}\},$$

则在估计类 $\mathcal{E}$ 中, $\hat{g}$ 的协方差阵对所有的 θ 一致地达到最小.

注意在定义估计类 $\mathcal{E}$ 时, 必须要求 δ 的各分量的方差为有限, 否则 $\mathrm{Cov}_\theta(\delta)$ 没有意义, 也就失去了比较的基础.

人们自然要问, 在估计类 $\mathcal{E}$ 中协方差阵最小能达到什么程度? 下面的多参数 C-R 不等式回答了这一问题.

定理 2.12 设样本 $X \sim f(x,\theta)d\mu(x)$, μ 为样本空间 $(\mathcal{X}, \mathcal{B})$ 上的一个 σ-有限测度, 参数 $\theta = (\theta_1, \cdots, \theta_k)^\tau \in \Theta$, Θ 为 $\mathbf{R}^k$ 中的一个开区域, $g(\theta) = \big(g_1(\theta), \cdots, g_r(\theta)\big)^\tau$, $r \leqslant k$. 假定

(1) 对 $\forall\, x \in \mathcal{X}, \theta \in \Theta$, $f(x,\theta) > 0$.

(2) 对 $\forall\, \theta \in \Theta$ 和 $1 \leqslant j \leqslant k$, $\partial f(x,\theta)/\partial\theta_j$ 存在, 且 $\int_{\mathcal{X}} \{\partial f(x,\theta)/\partial\theta_j\} d\mu(x) = 0$.

(3) 记 $S_j = S_j(x,\theta) = \partial \log f(x,\theta)/\partial\theta_j$. 对任一 $\theta \in \Theta$, 有

$$\mathrm{E}_\theta |S_i(X,\theta) S_j(X,\theta)| < \infty, \quad i,\ j = 1, \cdots, k, \quad \theta \in \Theta,$$

且方阵 $I(\theta) := \big(\mathrm{E}_\theta(S_i S_j), 1 \leqslant i, j \leqslant k\big)$ 是正定的.

在上述条件下, 若对估计 $\delta = (\delta_1, \cdots, \delta_r)^\tau \in \mathcal{E}$, 恒等式 $\int_{\mathcal{X}} \delta_i(x) f(x,\theta) d\mu(x) \equiv g_i(\theta)$ 可以在积分号下对 θ_j 取偏导数, $1 \leqslant i \leqslant r, 1 \leqslant j \leqslant k$, 则有

$$\mathrm{Cov}_\theta(\delta) \geqslant D(\theta) I^{-1}(\theta) D(\theta)^\tau, \ \theta \in \Theta, \tag{2.12}$$

其中 $D(\theta) = \partial g(\theta)/\partial\theta^\tau$ 为 $r \times k$ 矩阵, 其 (i,j) 元为 $\partial g_i(\theta)/\partial\theta_j$.

证明 记 $S = S(X,\theta) = (S_1(X,\theta), \cdots, S_k(X,\theta))^\tau$. 由假定 (2) 有 $\mathrm{E}_\theta[S(X,\theta)] = 0$, $\theta \in \Theta$. 由假定 (3), $I(\theta) = \mathrm{Cov}_\theta(S)$. 由加在 δ 上的条件知 $\mathrm{E}_\theta[\delta S(X,\theta)^\tau] =$

$\partial g(\theta)/\partial\theta^\tau = D(\theta)$. 故对任一 $r\times k$ 矩阵 C, 有

$$0 \leqslant \mathrm{Cov}_\theta(\delta - CS) = \mathrm{Cov}_\theta(\delta) - D(\theta)C^\tau - CD(\theta)^\tau + CI(\theta)C^\tau.$$

取 $C = D(\theta)I^{-1}(\theta)$, 即得 (2.12). ■

与单参数情形类似, $I(\theta)$ 称为有关分布族的 **Fisher 信息矩阵**. 它也具有与单参数情形 Fisher 信息函数类似的一些性质.

关于 (2.12) 中等号成立的条件, 此处不再讨论, 有兴趣的读者可参看陈希孺 (1981, 1997).

2.2 同变估计

2.2.1 同变性概论

估计的同变性 (equivariance) 要求, 是由相应的统计问题在有关的变换群之下的不变性决定的, 其中最简单也是最重要的例子, 就是估计量应与观测数据的测量原点及测量单位的改变保持同步. 设要估计的参数真值为 θ, 原先测得的样本为 $X_1,\cdots,X_n$, 用 $f(X_1,\cdots,X_n)$ 估计 θ. 现将测量原点移至 $-c$ 处, 则参数真值由 θ 变为 $\theta+c$, 样本由 $(X_1,\cdots,X_n)$ 变为 $(X_1+c,\cdots,X_n+c)$, 估计值由 $f(X_1,\cdots,X_n)$ 变为 $f(X_1+c,\cdots,X_n+c)$, 后者是 $\theta+c$ 的估计. 但在原先的原点之下, $\theta+c$ 这个量是用 $f(X_1,\cdots,X_n)+c$ 估计的. 这样, 从两种不同的度量原点出发, 我们对 $\theta+c$ 作出了两个估计, 从直觉上应要求两者相等, 即

$$f(X_1+c,\cdots,X_n+c) = f(X_1,\cdots,X_n)+c,\quad -\infty<c<\infty.$$

这一要求构成对所用的估计 f 的一个限制. 满足这一条件的估计称为在平移变换群下的同变估计 (equivariant estimates), 或位置同变估计. 类似地, 若考虑到随着测量单位的变化, 刻度参数和样本的测量值分别由 θ 和 $(X_1,\cdots,X_n)$ 变为 $c\theta$ 和 $(cX_1,\cdots,cX_n)$, 从而导致对估计量 f 的下述要求:

$$f(cX_1,\cdots,cX_n) = cf(X_1,\cdots,X_n),\quad c>0.$$

满足这一条件的估计称为在刻度变换群下的同变估计, 或刻度同变估计.

现在考虑一批产品的废品率 p 的估计问题, $0\leqslant p\leqslant 1$. 设独立随机地检查了 n 个产品, 令 $X_i=1$ 或 0, 视第 i 个为废品或合格品而定. 这样就得到了样本 $X_1,\cdots,X_n$. 设 用 a 估计 p 时引起的损失为 $L(p,a)=(p-a)^2$. 注意到可用变换 $x_i\mapsto 1-x_i, p\mapsto 1-p, a\mapsto 1-a$ 把估计废品率的问题转换为估计合格品率的等价问题, 此时损失函数不变, 自然希望估计量 δ 具有相应的性质, 即 $\delta(1-x_1,\cdots,1-x_n)=1-\delta(x_1,\cdots,x_n)$.

把上述情况加以推广, 就得到点估计理论中的同变性概念. 应当指出, 这一原理并不限于参数估计, 它也适用于其他统计决策问题. 本节涉及群和变换群的一些基本概念, 读者可参看任何一本抽象代数的教科书, 例如, 贾柯勃逊 (1960).

定义 2.4 (决策问题的不变性 (invariance of decision problem)) 设 $(\mathcal{X},\mathcal{B})$

为样本空间, $\mathcal{P}=\{P_\theta,\theta\in\Theta\}$ 为随机元 X 的分布族, 满足可识别性条件, 即

$$\text{若 } \theta_1\neq\theta_2,\ \text{则 } P_{\theta_1}\neq P_{\theta_2}. \tag{2.13}$$

(1) 假定在样本空间 $(\mathcal{X},\mathcal{B})$ 上存在一个可测变换群 G, 其中每个 $g\in G$ 都是 $\mathcal{X}$ 到自身上的一一变换, 且对任一 $B\in\mathcal{B}$, 有 $g^{-1}B\in\mathcal{B}$. 这后一要求等价于要求对 $\forall\ g\in G$, 有 $g(\mathcal{B})=\mathcal{B}$. **此时称样本空间在变换群 G 之下不变**.

(2) 假定对 $\forall\ g\in G$ 及 $\theta\in\Theta$, 当 $X\sim P_\theta$ 时, gX 的分布仍在分布族 $\mathcal{P}$ 中, 即存在 $\bar\theta\in\Theta$, 使

$$P_\theta(gX\in B)=P_\theta(X\in g^{-1}B)=P_{\bar\theta}(X\in B),\ B\in\mathcal{B}. \tag{2.14}$$

由可识别性条件 (2.13), 上述的 $\bar\theta$ 是由 g 唯一确定的, 因而可以定义空间 Θ 上的一个变换 $\bar g$: $\bar g\theta=\bar\theta$. 不难验证, 所有这样的 $\bar g$ 构成 Θ 上的一个变换群 $\bar G=\{\bar g:g\in G\}$. 为此首先证明, 对任意的 $g_1\in G,g_2\in G$, 有 $\bar g_1\bar g_2\in\bar G$, 且 $\bar g_1\bar g_2=\overline{g_1g_2}$. 这是因为, 若令 $g=g_1g_2$, 则按定义,

$$\begin{aligned}P_{\bar g\theta}(X\in B)&=P_\theta(X\in g^{-1}B)=P_\theta(X\in g_2^{-1}g_1^{-1}B)\\&=P_{\bar g_2\theta}(X\in g_1^{-1}B)=P_{\bar g_1\bar g_2(\theta)}(X\in B).\end{aligned}$$

由 (2.13) 即得 $\overline{g_1g_2}=\bar g_1\bar g_2$, 因而 $g\mapsto\bar g$ 即为 G 到 $\bar G$ 上保持乘法运算的一个映照, 不妨称之为一个同态. 作为群 G 的一个同态像, $\bar G$ 是一个群. 参看习题 2.14. 由此也附带证明了, 每个 $\bar g\in\bar G$ 都是 Θ 到自身上的一个一一变换. **此时样本分布族在变换群 $\bar G$ 之下不变**.

当以上两个条件都成立时, 称统计模型在变换群 G 之下不变.

(3) 设给定了决策空间 $(\mathcal{A},\mathcal{B}_\mathcal{A})$ 和定义于 $\Theta\times\mathcal{A}$ 上的损失函数 $L(\theta,a)$, 对每个 $\theta\in\Theta$, $L(\theta,\cdot)$ 为 $\mathcal{B}_\mathcal{A}$-可测. 我们假定对每个 $g\in G$, 存在 $(\mathcal{A},\mathcal{B}_\mathcal{A})$ 到自身上的一个一对一的可测变换 g^*, 使得

$$L(\bar g\theta,g^*a)=L(\theta,a),\ \text{对 }\forall\ \theta\in\Theta,\ a\in\mathcal{A}. \tag{2.15}$$

并且所有这样的 g^* 构成一个与 G 同态的群 $G^*=\{g^*:g\in G\}$. **与 $\bar G$ 不同的是, $\bar G$ 为群是可以证明的, 而 G^* 为群则是一个要求**.

如是, **若上述三个条件都成立, 则称以 $(\mathcal{X},\mathcal{B},\mathcal{P})$ 和 $L(\theta,a)$ 为标志的统计决策问题在变换群 G 之下不变**.

定义 2.5 (同变决策函数) 在上述统计架构之下, 一个随机化决策函数 $\delta(x,A)$, $x\in\mathcal{X},A\in\mathcal{B}_\mathcal{A}$, 称为变换群 G 之下的一个**同变决策函数**, 如果

$$\delta(gx,g^*A)=\delta(x,A),\ \text{对 }\forall g\in G,x\in\mathcal{X},A\in\mathcal{B}_\mathcal{A}. \tag{2.16}$$

一个非随机化决策函数 $\delta(\cdot)$ 称为变换群 G 之下的一个同变决策函数, 如果

$$\delta(gx)=g^*\delta(x),\ \text{对 }\forall g\in G,x\in\mathcal{X}. \tag{2.17}$$

而在一个不变的参数估计问题中, 满足上述同变性要求的估计则称为**同变估计**. 显然, 非随机化同变估计类是随机化同变估计类的一个子类.

今设在同变估计的架构下, 我们用一同变估计 $\delta(X)$ 去估计某一参数函数 $h(\theta)$. 由同变性的定义, 对任一 $g\in G$, $\delta(gX)=g^*\delta(X)$. 考虑到 $X\sim P_\theta$ 时 $gX\sim P_{\bar g\theta}$,

$\delta(gX)$ 既可用于估计 $h(\bar{g}\theta)$, 又可用于估计 $g^*h(\theta)$. 因此, 我们可以顺理成章地要求

$$h(\bar{g}\theta) = g^*h(\theta), \text{ 对 } \forall\ g \in G, \theta \in \Theta, \tag{2.18}$$

因而 $h(\bar{g}\theta)$ 的取值仅与 $h(\theta)$ 有关. 对参数函数的这一要求是与估计的同变性要求相匹配的.

由 (2.14), 即 $P_\theta(gX \in B) = P_{\bar{g}\theta}(X \in B),\ B \in \mathcal{B}$, 可以看出, 对任一非负 $\mathcal{B}$-可测函数 ψ, 有

$$\mathrm{E}_\theta\big[\psi(gX)\big] = \mathrm{E}_{\bar{g}\theta}\big[\psi(X)\big], \text{ 对 } \forall g \in G. \tag{2.19}$$

这一公式在后面有用.

例 2.6 设 $X_1, \cdots, X_n$ 是从分布 $F(\cdot - \theta)$ 中抽取的 i.i.d. 样本, 其中 F 为已知的一维分布函数, $\theta \in \Theta = (-\infty, \infty)$, 要估计参数 θ. 损失函数为 $L(\theta, a)$. 考虑样本空间 $(\mathbf{R}^n, \mathcal{B}_n)$ 上的平移群 $G = \{g_c : -\infty < c < \infty\}$, 其中

$$g_c(x) = x + c\mathbb{1}_n,\ x = (x_1, \cdots, x_n)^\tau, \mathbb{1}_n = (1, \cdots, 1)^\tau \in \mathbf{R}^n. \tag{2.20}$$

这个变换在 Θ 中导出的变换为 $\bar{g}_c(\theta) = \theta + c$, 因为若 $X_1 \sim F(x - \theta)$, 则 $X_1 + c \sim F(x - \theta - c)$. $\bar{G} = \{\bar{g}_c : -\infty < c < \infty\}$ 构成一个群. 取 $G^* = \{g_c^* : -\infty < c < \infty\}$, 其中 $g_c^*(a) = a + c,\ a \in \mathcal{A} = (-\infty, \infty)$. 易见, 为要这一估计问题在变换群 G 之下不变, 必须而且只需将损失函数取为 $L(\theta, a) = W(a - \theta)$ 的形式. 此时, θ 的同变估计 $\delta(X)$ 是指满足条件

$$\delta(x + c\mathbb{1}_n) = \delta(x) + c,\ -\infty < c < \infty$$

的任一估计. 要求同变性限制了所允许考虑的估计的范围, 凡不满足上述同变性要求的估计量都被排除在外. 在某些情况下, 在这缩小了的估计类中, 就可以找到在某种准则下的最优估计. 在这个意义上说, 这和引进无偏性的要求出于相同的考虑.

2.2.2 风险一致最小的同变估计

我们研究同变估计理论, 自然希望找到最好的同变估计. 为此, 我们引进下面的定义.

定义 2.6 设 δ 是在变换群 G 之下不变的估计问题中的一个同变估计, 其风险函数记为 $R(\theta, \delta)$, 如果对于任一同变估计 δ^*, 总有

$$R(\theta, \delta) \leqslant R(\theta, \delta^*), \text{ 对 } \forall\ \theta \in \Theta,$$

则称 δ 为**风险一致最小的同变估计** (minimum risk equivariant estimate), 简称**最小风险同变 (MRE) 估计**, 或**最优同变估计**.

以上我们讲到, 引进同变性的作用在于缩小了所考虑的估计的范围, 有利于寻求最优估计. 其实更重要的是, 同变性的要求简化了风险函数, 而决策函数的优劣取决于它们的风险函数.

为说明同变性如何简化了风险函数, 我们引进以下概念. 设 $\mathcal{G}$ 为空间 Ω 上的一个变换群. 在 Ω 中任取两点 $\omega_1,\ \omega_2$, 若存在 $g \in \mathcal{G}$, 使 $\omega_2 = g(\omega_1)$, 则称这两点是

等价的. 于是空间 Ω 被分成一些两两互斥的等价类, 每个等价类称为 $\mathcal{G}$ 的一个**轨道** (orbit). 易见, 每个 ω 所在的轨道即为 Ω 的子集 $\{g(\omega): g\in\mathcal{G}\}$. 如果 $\mathcal{G}$ 只有一个轨道, 则称 $\mathcal{G}$ 在 Ω 上是**可递的**. 我们有如下定理.

定理 2.13 *设 δ 为在变换群 G 之下不变的估计问题中的一个同变估计, 则 δ 的风险函数在 $\bar{G}$ 的每个给定的轨道上取常数值, 亦即其风险函数满足*

$$R(\bar{g}\theta,\delta)=R(\theta,\delta),\ \text{对}\ \forall\ \theta\in\Theta,\bar{g}\in\bar{G}.$$

证明 为使读者容易理解随机化同变估计情形的证明, 先设 δ 为一非随机化同变估计. 依次利用 (2.19), (2.17) 和 (2.15), 立得

$$\mathrm{E}_{\bar{g}\theta}[L\big(\bar{g}\theta,\delta(X)\big)]=\mathrm{E}_{\theta}[L\big(\bar{g}\theta,\delta(gX)\big)]=\mathrm{E}_{\theta}[L\big(\bar{g}\theta,g^{*}\delta(X)\big)]=\mathrm{E}_{\theta}[L\big(\theta,\delta(X)\big)],$$

即 $R(\bar{g}\theta,\delta)=R(\theta,\delta)$.

今设 $\delta(X,da)$ 为一随机化同变估计, 则由 $\mathrm{E}_{\bar{g}\theta}[\big(\psi(X)\big)]=\mathrm{E}_{\theta}\big[\psi(gX)\big]$, 有

$$\begin{aligned}\mathrm{E}_{\bar{g}\theta}\left[\int_{\mathcal{A}}L(\bar{g}\theta,a)\delta(X,da)\right]&=\mathrm{E}_{\theta}\left[\int_{\mathcal{A}}L(\bar{g}\theta,a)\delta(gX,da)\right]\\&=\mathrm{E}_{\theta}\left[\int_{\mathcal{A}}L(\bar{g}\theta,g^{*}a)\delta(gX,dg^{*}a)\right].\end{aligned}$$

第二个等式用到 g^{*} 为 $\mathcal{A}$ 到自身上的一一变换. 再利用 (2.16) 和 (2.15), 即 $\delta(gx,dg^{*}a)=\delta(x,da)$ 和 $L(\bar{g}\theta,g^{*}a)=L(\theta,a)$, 即得

$$R(\bar{g}\theta,\delta)=\mathrm{E}_{\theta}\left[\int_{\mathcal{A}}L(\theta,a)\delta(X,da)\right]=R(\theta,\delta).\qquad\blacksquare$$

由于有了上述定理, 我们也可以说, 同变性的要求缩小了参数空间, 即在 $\bar{G}$ 的每条轨道上只要取一个点就够了. 例如, 在平移群 (2.20) 之下, 由形如 $\bar{g}_{c}(\theta)=\theta+c$ $(-\infty<c<\infty)$ 的变换组成的变换群 $\bar{G}$ 只有一条轨道, 参数空间实质上被压缩为一个点. 因而在一些简单情况下, 最小风险同变估计存在.

为寻求最优同变估计, 首先我们要问, 如果采用凸损失, 是否所有非随机化决策函数仍然构成一个本质完全类. 设 $\delta(x,A),x\in\mathcal{X},A\in\mathcal{B}_{\mathcal{A}}$ 为一随机化同变估计. 基于 δ 可构作一非随机化估计 $\delta_{1}(X)$,

$$\delta_{1}(x)=\int_{\mathcal{A}}a\delta(x,da).$$

此处需要假定 $\mathcal{A}$ 为 $\mathbf{R}^{k}$ 中的开凸集, 且对 $\forall\ x\in\mathcal{X}$, $\delta_{1}(x)\in\mathcal{A}$. 与 2.1.1 节类似, 由 Jensen 不等式, 有 $\mathrm{E}_{\theta}[L\big(\theta,\int_{\mathcal{A}}a\delta(X,da)\big)]\leqslant\mathrm{E}_{\theta}[\int_{\mathcal{A}}L(\theta,a)\delta(X,da)]$, 即

$$R(\theta,\delta_{1})\leqslant R(\theta,\delta),\ \text{对}\ \forall\ \theta\ \in\Theta.$$

问题在于 δ_{1} 能否满足同变性要求. 为简单起见, 设 $\mathcal{A}=\mathbf{R}^{k}$, 则由 $\delta(gx,dg^{*}a)=\delta(x,da)$, 有

$$\delta_{1}(gx)=\int_{\mathcal{A}}a\delta(gx,da)=\int_{\mathcal{A}}g^{*}a\delta(gx,dg^{*}a)=\int_{\mathcal{A}}g^{*}a\delta(x,da).$$

"同变性" 要求对任一 g 和 x, $g^{*}\delta_{1}(x)=\delta_{1}(gx)$, 即 $g^{*}\int_{\mathcal{A}}a\delta(x,da)=\int_{\mathcal{A}}g^{*}a\delta(x,da)$. 在一般情况下, 这一要求无法满足; 而当 G^{*} 为 $\mathbf{R}^{k}$ 上的线性变换群 (又称仿射群) 时

上述条件自然满足.

下面即以一些简单情形为例说明如何寻求最小风险同变估计. 由于涉及的变换群是仿射群, 在凸损失之下, 寻求最优同变估计只需限于非随机化估计的类中.

例 2.7 设 $X_1,\cdots,X_n$ 是从分布 $N(\theta,1)$ 中抽取的 i.i.d. 样本, $\theta\in\Theta=(-\infty,\infty)$, 要估计 θ, 取平方损失 $L(\theta,a)=(\theta-a)^2$. 求在平移群 (2.20) 之下 θ 的最优同变估计, 即 MRE 估计.

首先, 我们引进变换群下不变量的概念. 记 $x=(x_1,\cdots,x_n)^\tau$. 样本空间上的函数 $u(x)$ 称为**变换群 G 下的一个不变量**, 如果它在群 G 的每条轨道上取常数值. 如果某一不变量 $t(x)$ 在不同轨道上总是取不同的值, 则 $t(x)$ 称为该变换群下的一个**极大不变量**. 易见, 若 $t(x)$ 为极大不变量, 则 $u(x)$ 为一不变量的充要条件为, $u(x)$ 可表示为 $t(x)$ 的函数.

现在我们找出同变估计的一般形式. 为此, 取定一特定的同变估计 $\delta_0(X)$, 易见任一估计 $\delta(X)$ 为一同变估计的充要条件为, 函数 $u(x)=\delta(x)-\delta_0(x)$ 对任何实数 c 满足 $u(g_cx)=u(x)$, 亦即 $u(x)$ 为平移群 (2.20) 下的一个不变量.

容易验证, 对平移群 (2.20) 而言, 一个极大不变量为 $y=(x_2-x_1,\cdots,x_n-x_1)^\tau$. 取 $\delta_0(x)=\bar{x}$, 即得 θ 的同变估计的一般形式为

$$\delta(X)=\bar{X}+v(Y),Y=(X_2-X_1,\cdots,X_n-X_1)^\tau.$$

要定出 v, 使

$$R(\theta,\delta)=\mathrm{E}_\theta\big(\delta(X)-\theta\big)^2=\mathrm{E}_\theta\big(\bar{X}-\theta+v(Y)\big)^2$$

达到最小. 因为 $(\bar{X},Y^\tau)$ 的联合分布为 n-维正态分布, 且 $\mathrm{E}_\theta[(\bar{X}-\theta)Y]=0$, 知 $\bar{X}$ 与 Y 相互独立, 故 $\bar{X}-\theta$ 与 $v(Y)$ 独立, 因而

$$R(\theta,\delta)=\mathrm{E}_\theta(\bar{X}-\theta)^2+\mathrm{E}_\theta[v^2(Y)].$$

上式在 $v\equiv0$ 时达到最小. 由此得出, 样本均值 $\bar{X}$ 即为在平移群 (2.20) 之下 θ 的 MRE 估计.

注 2.1 在该例中, $\bar{X}$ 是充分统计量, 由充分性原则自然会联想到, 似乎可以只考虑基于 $\bar{X}$ 的同变估计, 而这种估计必有 $\bar{X}+c$ 的形式, 其中又只有在 $c=0$ 时风险最小, 由此立得 $\bar{X}$ 为最优同变估计. 但问题在于, 加上同变性限制后充分性原则是否仍然成立. 已经证明, 在很一般的条件下这一结论成立. 参看陈桂景, 陈希孺 (1982).

例 2.8 设 n-维随机向量 $X=(X_1,\cdots,X_n)^\tau\sim F(x-\theta\mathbb{1}_n)$, 此处 F 已知, $\mathbb{1}_n=(1,\cdots,1)^\tau\in\mathbf{R}^n$, $\theta\in\Theta=(-\infty,\infty)$. 损失函数取为 $L(\theta,a)=(\theta-a)^2$. 假定 $\mathrm{E}_0(X_1^2)<\infty$, 此处的期望是在参数 $\theta=0$ 之下取的. 求在平移群 (2.20) 之下 θ 的 MRE 估计.

前面已指出, 在平移群 (2.20) 之下, 参数空间上的变换群只有一条轨道, 故可固定 $\theta=0$ 来讨论. 由于 $\delta_0(X)=X_1$ 为一同变估计, 可知 θ 的同变估计的一般形式为

$$\delta(X)=X_1-v(Y),Y=(X_2-X_1,\cdots,X_n-X_1)^\tau:=(Y_2,\cdots,Y_n)^\tau.$$

问题归结为找出 v, 使 $R(\theta,\delta)=\mathrm{E}_0\big(X_1-v(Y)\big)^2$ 达到最小. 按条件期望的性质, 有

$$R(\theta,\delta)=\mathrm{E}_0\{\mathrm{E}_0[(X_1-v(Y))^2|Y]\}.$$

易见, 条件期望在 $v(Y)=\mathrm{E}_0(X_1|Y)$ 时取得最小值, 故所求 MRE 估计即为

$$\delta(X)=X_1-\mathrm{E}_0(X_1|Y). \tag{2.21}$$

今设 X 有密度函数 $f(x-\theta\mathbb{1}_n)$, f 已知. 在 $\theta=0$ 时, X 有密度 $f(x)=f(x_1,\cdots,x_n)$. 作线性变换$(X_1,\cdots,X_n)\mapsto(X_1,Y^\tau):=(X_1,X_2-X_1,\cdots,X_n-X_1)$, 知 (X_1,Y^τ) 的联合密度为 $f(x_1,y_2+x_1,\cdots,y_n+x_1)$. 给定 $Y=y$, X_1 的条件密度为

$$\frac{f(x_1,y_2+x_1,\cdots,y_n+x_1)}{\int f(s,y_2+s,\cdots,y_n+s)ds}.$$

故

$$\begin{aligned}\delta(x)&=x_1-\frac{\int sf(s,y_2+s,\cdots,y_n+s)ds}{\int f(s,y_2+s,\cdots,y_n+s)ds}\\&=x_1-\frac{\int(x_1-t)f(x-t\mathbb{1}_n)dt}{\int f(x-t\mathbb{1}_n)dt}=\frac{\int tf(x-t\mathbb{1}_n)dt}{\int f(x-t\mathbb{1}_n)dt}.\end{aligned} \tag{2.22}$$

上述估计是 Pitman 在 1939 年引进的, 一般称之为 **Pitman 估计**.

例 2.9 (刻度参数族) 设 $X_1,\cdots,X_n$ 具有联合密度

$$\theta^{-n}f(x/\theta),\quad x=(x_1,\cdots,x_n)^\tau,\quad \theta>0,$$

其中 f 为一已知函数, θ 为一未知的刻度参数. 这个模型在形如

$$g_c(x)=cx,\quad \bar g_c(\theta)=c\theta,\quad c>0 \tag{2.23}$$

的变换之下是不变的.

假定我们要估计的是 $h(\theta)=\theta^r$. 前面已指出, 为了估计的同变性, 要求对任何 g 和 θ, $h(\bar g_c\theta)=g_c^*h(\theta)$, 即

$$g_c^*a=c^ra,\quad a\in\mathcal{A}=(0,\infty),$$

此处 $\mathcal{A}$ 为行动空间. 损失函数在上述变换下不变的条件是

$$L(\theta,\ a)=L(c\theta,\ c^ra)=L(1,a/\theta^r):=\gamma(a/\theta^r). \tag{2.24}$$

θ^r 的估计 $\delta(X)$ 为刻度同变的条件是

$$\delta(cX)=c^r\delta(X). \tag{2.25}$$

因为变换群 $G=\{\bar g_c,c>0\}$ 在参数空间 $(0,\infty)$ 上是可递的, 由定理 2.13, 任何同变估计的风险为一常数. 因此我们可以相信 MRE 估计是存在的.

为了导出 θ^r 的 MRE 估计, 先设 $\delta_0(X)$ 为一同变估计. 因为在刻度变换群下,

$$z=(z_1,z_2,\cdots,z_n):=(x_1/|x_1|,x_2/x_1,\cdots,x_n/x_1)$$

为一极大不变量, 因而 $\delta(X)$ 为一同变估计的充要条件是存在一个函数 $w(z)$, 使得

$$\delta(X)=\delta_0(X)/w(Z).$$

以 E_1 记当参数 $\theta=1$ 时的期望算子. 今设存在 θ^r 的一个风险有限的同变估计 $\delta_0(X)$, 且对每个 z, 存在一个数 $w(z)=w^*(z)$, 使

$$\mathrm{E}_1[\gamma\big(\delta_0(X)/w(z)\big)|\ z] \tag{2.26}$$

达到最小, 则 θ^r 的 MRE 估计即为

$$\delta^*(X)=\delta_0(X)/w^*(Z). \tag{2.27}$$

读者不难给出上述结论的证明, 也可参看 Lehmann 和 Casella (1998) 第三章定理 3.1 和 3.3.

例 2.10 设 $X_1,\cdots,X_n$ 是取自分布 $F\big((\cdot-\theta_1)/\theta_2\big)$ 的 i.i.d. 样本, $n\geqslant 2$, 分布函数 F 已知, $\theta_1\in\mathbf{R},\theta_2>0,\theta=(\theta_1,\theta_2)$, 要估计 (θ_1,θ_2^r), 损失函数为 $L(\theta,a)=W\big((\theta_1-a_1)/\theta_2,a_2/\theta_2^r\big)$. 此处的参数 θ_1 和 θ_2 分别称为位置(或转移)参数和刻度参数, 而 $a=(a_1,a_2)$. 本例考虑的分布族即为位置 – 刻度族.

样本空间 $(\mathbf{R}^n,\mathcal{B}_n)$ 上的变换群 G 可取为由所有形如

$$g_{c,d}(x)=cx+d\mathbb{1}_n,\quad c>0,-\infty<d<\infty$$

的变换组成的仿射群. 参数空间上相应的变换群 $\bar{G}$ 由所有形如

$$\bar{g}_{c,d}(\theta_1,\theta_2)=(c\theta_1+d,c\theta_2),\quad c>0,-\infty<d<\infty$$

的变换组成. 行动空间上的变换群 G^* 取成由所有形如

$$g^*_{c,d}(a_1,a_2)=(ca_1+d,c^ra_2),\quad c>0,-\infty<d<\infty$$

的变换组成的变换群. 易见估计问题在所给的损失之下是不变的, 而同变估计 $\delta=(\delta_1,\delta_2)$ 满足的条件为

$$\delta_1(cx+d\mathbb{1}_n)=c\delta_1(x)+d, \tag{2.28}$$

$$\delta_2(cx+d\mathbb{1}_n)=c^r\delta_2(x). \tag{2.29}$$

变换群 $\bar{G}$ 为参数空间上的一个线性可递群, 它只有一条轨道, 因而可以期望, 当 W 取为二元凸函数时, 最优同变估计存在, 其形式取决于分布 F. 下面考虑一种最重要的情况, 即 F 为正态分布 $N(0,1)$, 且损失函数形如

$$L(\theta,a)=\lambda_1(a_1-\theta_1)^2/\theta_2^2+\lambda_2(a_2-\theta_2^r)^2/\theta_2^{2r}, \tag{2.30}$$

其中 $\lambda_1>0,\lambda_2>0$ 已知. 此时 (θ_1,θ_2^r) 的 MRE 估计为

$$\hat{\delta}_1(X)=\bar{X},\ \hat{\delta}_2(X)=\Big[\frac{\Gamma\big((n-1+r)/2\big)}{2^{r/2}\Gamma\big((n-1+2r)/2\big)}\Big]S^r,\ S^2=\sum_{i=1}^n(X_i-\bar{X})^2. \tag{2.31}$$

证明步骤如下: 设 $\big(\delta_1(X),\delta_2(X)\big)$ 为 (θ_1,θ_2^r) 的一个同变估计.

1° 在 (2.28) 中令 $c=1$, d 任意, 知 δ_1 符合例 2.7 中对同变估计的要求, 按损失函数 (2.30) 的形式, 由例 2.7 得出 θ_1 的 MRE 估计为 $\hat{\delta}_1(X)=\bar{X}$.

2° 在 (2.29) 中令 $c=1$, 知 δ_2 为平移群下的一个不变量, 故 $\delta_2(x)=u(y):=u(x_2-x_1,\cdots,x_n-x_1)$. 再在 (2.29) 中令 $d=0$, 知 u 对任一正数 c 满足条件 $u(cy)=c^ru(y),y\in\mathbf{R}^{n-1}$. 注意到 S 可表示为 $y:=(x_2-x_1,\cdots,x_n-x_1)$ 的函数 $S(y)$, $u(y)/S^r(y)$ 在刻度变换群下是不变的. 由例 2.9, 知 $\mathbf{R}^m$ 上的函数 $\zeta(y_1,\cdots,y_m)$ 在刻度变换群下不变, 当且仅当 $\zeta(y_1,\cdots,y_m)=\gamma(y_1/|y_1|,y_2/y_1,\cdots,y_m/y_1)$. 因此存在

$$Z=v\Big(\frac{X_2-X_1}{|X_2-X_1|},\frac{X_3-X_1}{X_2-X_1},\cdots,\frac{X_n-X_1}{X_2-X_1}\Big),$$

使 $\delta_2(X) = S^r Z$. 此即 θ_2^r 的同变估计的一般形式.

3°. 因为 $\bar{G}$ 在放射群下只有一个轨道, 由定理 2.13, 同变估计 $S^r Z$ 的风险函数等于常数, 故只需在 $N(0,1)$ 的分布下计算其风险. 由 Basu 定理, 知 S 与 Z 独立, 故

$$\mathrm{E}_{(0,1)}[(S^r Z - 1)^2 | z] = \mathrm{E}_{(0,1)}(S^r z - 1)^2,$$

上式当 $z = z^* := \mathrm{E}_{(0,1)}(S^r)/\mathrm{E}_{(0,1)}(S^{2r})$ 时取得最小值, 故 θ_2^r 的 MRE 估计即为 $\hat{\delta}_2(X) = z^* S^r$. 因为在 $N(0,1)$ 之下, $S^2 \sim \chi^2_{n-1}$, 通过简单的计算, 即可得到 (2.31).

在变换群 $\bar{G}$ 有多条轨道的情形, 最优同变估计常常不存在. 但不可一概而论. 以本节开头估计废品率 p 的问题为例, 对任一 $p \in [0, 1/2]$, 集合 $\{p, 1-p\}$ 均为变换群 $\bar{G}$ 的轨道, 但样本均值 $\bar{X}$ 即为最优同变估计. 这一论断立即可由充分性原则推得.

在历史上, 位置和刻度参数的同变估计的理论最早见于 Pitman (1939), 同变估计的一般理论最先由 Peisakoff (1950) 和 Kiefer (1957) 提出. 本节只涉及同变估计的一些简单的讨论, 读者欲知其详, 可参看 Lehmann (1983), 陈希孺 (1981, 1997) (1999, 2009), 以及成平, 陈希孺, 陈桂景, 吴传义 (1985).

2.3 Bayes 估计

在统计决策理论的框架下, 风险函数一致最优准则涉及的往往是无穷多个指标下的优化问题, 一般没有最优解. 前两节的主题是, 引进某种要求以限制入围的估计量, 并在缩小了的估计类中, 找出可能存在的一致最优估计. 下面要讲的 Bayes 估计和 minimax 估计则是循着另一种思路, 即不是限制估计量的类, 而是放宽风险函数 "一致最优" 的评判准则, 而代之以一种整体性的指标. 因为此时涉及的只是单个指标之下的优化问题, 最优解一般都存在. 本节首先主要介绍统计决策框架下的 Bayes 估计, 然后简单讨论与 Bayes 统计推断有关的一些问题.

2.3.1 Bayes 统计决策的基本框架

本小节介绍 Bayes 决策的基本架构. 为此, 设样本空间和样本分布族分别为 $(\mathcal{X}, \mathcal{B})$ 和 $\{P_\theta,\ \theta \in \Theta\}$, X 为取自该分布族中的一个样本. 在参数空间 Θ 和行动空间 $\mathcal{A}$ 上分别引进由各自的子集构成的 σ-域 $\mathcal{B}_\Theta$ 和 $\mathcal{B}_\mathcal{A}$. 在估计问题中, 通常将这些空间取为欧氏的. 在涉及 Bayes 风险分析的问题中, 损失函数 $L(\theta, a)$ 总是假定为 $\mathcal{B}_\Theta \times \mathcal{B}_\mathcal{A}$-可测函数. 应当指出, 在决策理论的一般框架下, 没有必要指出有关问题是参数估计、假设检验, 还是别的什么, 问题的性质体现在决策空间 $(\mathcal{A}, \mathcal{B}_\mathcal{A})$ 和损失函数 $L(\theta, a)$ 的表述之中. 在后面的行文中我们常以 "估计" 称之, 且经常省去被估函数 $g(\theta)$ 不提. 另一方面, 讨论中或多或少都会涉及 Bayes 学派的一些观点.

作为现代统计中很有影响的一个学派, Bayes 学派可以溯源到 Bayes 1763 年的基础论文. Laplace 大大推广了 Bayes 方法, 特别是使用了无信息先验, 他的贡献体

现在他 1774 年的论文和 1820 年的重要著作中. 但他也采用非 Bayes 方法, 始终没有对两者作出明显的区分. Jeffreys (1939 第 1 版, 1961 第 3 版) 系统论述了基于无信息先验的统计推断的理论, 推广和发展了 Laplace 的方法. de Finetti (1937) (1970) 和 Savage (1954) 对主观理论的发展, Raiffa 和 Schlaifer (1961) 对共轭先验的使用都起了重要作用.

定义 2.7 首先, 在 $\mathcal{B}_\Theta$ 上引进一个概率分布 ξ , 称为 θ 的**先验分布** (prior distribution). 设 δ 为一决策函数, 其风险函数为 $R(\theta,\delta)$. 称

$$R_\xi(\delta)=\int_\Theta R(\theta,\delta)d\xi(\theta) \tag{2.32}$$

为 δ 关于先验分布 ξ 的 **Bayes 风险** (Bayes risk). 如果某个决策函数 δ_ξ 满足

$$R_\xi(\delta_\xi)\leqslant R_\xi(\delta),\ \text{对所有 } \delta, \tag{2.33}$$

则称 δ_ξ 为所提问题在先验分布 ξ 之下的 **Bayes 解**. 在点估计问题中, 称为 **Bayes 估计**. 在上述架构之下的统计决策问题称为 **Bayes 决策问题**.

在上述定义中, (2.33) 中 "对所有 δ" 成立自然也可以局限为对满足某种条件的 δ 成立.

此处与以往不同的是先验分布的引进. 按 Bayes 学派的观点, 即使在进行观测以得到样本 x 之前, 人们对 θ 也会有一些知识. 这些知识可以来自于某种理论、以往的经验, 或者研究者的主观认识, 并且可以用一个先验分布来表达. 这样的先验分布在一种情况下易于为频率学派所接受, 即假定有关的统计问题不是一次性的, 而是经常重复出现的; 每次参数 θ 的取值不同, 其出现频率遵从某个概率分布 ξ. 例如逐日从生产线上抽取若干产品估计当日的废品率 θ, 每天的 θ 都会有波动, 可以认为逐日的 θ 取值服从某个概率分布. 在这种情况下, θ 可以视为一个随机变量. 在多数情形 θ 无 "随机" 可言. 例如当 θ 为需要确定的某个未知的物理常数时, 关于 θ 的概率陈述可解释为一种主观概率, 它反映个人对于这种陈述的似然性的相信程度. Bayes 学派认为, 主观的方法可以应用到缺少甚至没有经验资料的情形. 在后一种情形, 先验分布也刻画了关于 θ 无知的程度. 而当获取样本 x 之后, 即可利用 θ 的后验分布, 即在 $X=x$ 给定时的条件分布, 对先验知识加以修正. 读者若要详细了解, 可参阅 Savage (1954), Box 和 Tiao (1973), Berger (1985) 等.

Bayes 方法的意义还在于它可以作为一种数学工具. Wald 决策理论中的一个重要结论是, 在一定条件下, Bayes 解及其在一定意义下的极限全体构成了全体决策函数的一个完全类. 因此, 它在统计决策理论中起着重要作用. 在处理容许性和求解 minimax 问题时, Bayes 方法也不失为一种有用的方法. 另外, 在有些问题中, 对有关估计的取值范围有一定的限制, 例如需要对某个统计量的方差构造一个非负估计, 借助于先验分布不难构造这样的估计. 对频率学派来讲, 如前所述, 在具有丰富的经验资料的情况下, 先验分布可以有一种频率解释, 这些信息的利用无疑有助于得出更为有效的估计. 即使有关统计问题只是一次性的, 或者虽有多次重复, 分布

ξ 也无法知道, 此时我们仍可以回到原来的出发点, 把 $R_\xi(\delta)$ 看成便于比较的一种综合性指标, 此时我们甚至可将 ξ 取为 $(\Theta, \mathcal{B}_\Theta)$ 上的一个无穷测度.

回到 ξ 为 θ 的先验分布的假定. 在这种架构下, 若将 (X, θ) 视为一个随机元, 则其联合分布为 $P(dx, d\theta) := P_\theta(dx)\xi(d\theta)$. 将 X 在 P 之下的**边缘分布**记为 P^X, 有

$$P^X(B) := P(X \in B) = \int_\Theta P_\theta(B) d\xi(\theta), \quad B \in \mathcal{B}.$$

原先的样本分布 P_θ, 则成为分布 P 之下给定 θ 时 X 的条件分布.

在 Bayes 分析的理论体系中, 后验分布起着关键的作用.

定义 2.8 设 ξ 为 θ 的给定的先验分布, 则在样本 $X = x$ 给定的条件下, θ 的条件分布称为 θ 的**后验分布** (posterior distribution), 记为 $P(\cdot|x)$.

有关后验分布的计算一般是很难的. 但在样本分布族 $\{P_\theta, \theta \in \Theta\}$ 受控于 $\mathcal{B}$ 上一个 σ-有限测度 μ 的重要情形, 问题不难解决. 此时若记 $dP_\theta(x) = f(x, \theta)d\mu(x)$, 则由条件概率的定义, 有

$$P(d\theta|x) = \frac{f(x,\theta)\, d\xi(\theta)}{\int_\Theta f(x,\theta)\, d\xi(\theta)}. \tag{2.34}$$

一般地说, 只要 $(\mathcal{X} \times \Theta, \mathcal{B} \times \mathcal{B}_\Theta)$ 是欧氏的, 则如定理 1.7 所示, 给定 x 时 θ 的条件概率分布存在, 这保证了 $P(d\theta|x)$ 有意义. 但除了 (2.34) 的情况之外, 没有合适的表达式.

由条件期望的性质, 当 δ 为非随机化估计时, 其 Bayes 风险可写为

$$\begin{aligned} R_\xi(\delta) &= \mathrm{E}[L(\theta, \delta(X)] = \mathrm{E}\{\mathrm{E}[L(\theta, \delta(X))|X]\} \\ &= \int_{\mathcal{X}} dP^X(x) \int_\Theta L(\theta, \delta(x)) P(d\theta|x). \end{aligned} \tag{2.35}$$

当 δ 为随机化估计时, 其 Bayes 风险可写为

$$\begin{aligned} R_\xi(\delta) &= \int_\Theta d\xi(\theta) \int_{\mathcal{X}} dP_\theta(x) \int_{\mathcal{A}} L(\theta, a)\delta(x, da) \\ &= \int_{\mathcal{X}} dP^X(x) \int_{\mathcal{A}} \delta(x, da) \int_\Theta L(\theta, a) P(d\theta|x). \end{aligned} \tag{2.36}$$

后验风险最小的原则

在后面我们将会看到, 在很普遍的情况下, Bayes 解都是非随机化的.

定义 2.9 表达式

$$R_\xi(\delta, x) = \begin{cases} \displaystyle\int_\Theta L(\theta, \delta(x)) P(d\theta|x), & (\delta \text{ 为非随机化}) \\ \displaystyle\int_{\mathcal{A}} \delta(x, da) \int_\Theta L(\theta, a) P(d\theta|x), & (\delta \text{ 为随机化}) \end{cases} \tag{2.37}$$

称为 δ 的**后验风险** (posterior risk).

无论何种情形, 都有

$$R_\xi(\delta) = \int_{\mathcal{X}} R_\xi(\delta, x) dP^X(x). \tag{2.38}$$

由此不难看出, 如果存在使后验风险达到最小的决策函数, 则它即为上述决策问题的 Bayes 解. 确切地说, 我们有下述基本定理.

定理 2.14 设 δ_0 为一非随机化决策函数, 且存在 $N \in \mathcal{B}, P^X(N) = 0$, 使得当 $x \notin N$ 时, 极值

$$\inf_{a\in\mathcal{A}} \int_\Theta L(\theta, a)P(d\theta|x) \tag{2.39}$$

在 $a = \delta_0(x)$ 处达到, 则在所有决策函数的类中, δ_0 是有关决策问题的一个 Bayes 解.

证明 设 δ 为任一随机化决策函数, 则由 (2.37), 当 $x \notin N$ 时,

$$R_\xi(\delta, x) = \int_{\mathcal{A}} \delta(x, da) \int_\Theta L(\theta, a)P(d\theta|x) \geqslant \int_\Theta L(\theta, \delta_0(x))P(d\theta|x) = R_\xi(\delta_0, x).$$

因而由 (2.38) 及 $P^X(N) = 0$, 即得 $R_\xi(\delta) \geqslant R_\xi(\delta_0)$. 证毕. ■

这个定理显示一个重要事实, 即在求 Bayes 解时, 可以只考虑非随机化决策函数. 反之, 若 δ_0 为一非随机化 Bayes 解, 其 Bayes 风险有限, 则在一定的条件下, 存在 $N \in \mathcal{B}, P^X(N) = 0$, 使得当 $x \notin N$ 时, 极值 (2.39) 在 $a = \delta_0(x)$ 处达到. 事实上, 若存在 $B \in \mathcal{B}, P^X(B) > 0$, 使得当 $x \in B$ 时, 极值 (2.39) 不能在 $\delta_0(x)$ 处达到, 则对 $x \in B$, 可找到 $a(x) \in \mathcal{A}$, 使 $\int_\Theta L(\theta, \delta_0(x))P(d\theta|x) > \int_\Theta L(\theta, a(x))P(d\theta|x)$. 定义 $\delta(x) = a(x)$ 或 $\delta_0(x)$, 视 $x \in B$ 是否成立而定. 若由此定义的 δ 为 $\mathcal{B}|\mathcal{B}_\mathcal{A}$ 可测, 则由 $P^X(B) > 0$, 容易证明 $R_\xi(\delta_0) > R_\xi(\delta)$, 与 δ_0 为 Bayes 解的假定相矛盾.

上述定理告诉我们, 为寻求 Bayes 估计, 只需在获得样本 x 之后, 找出使 $\int_\Theta L(\theta, a)P(d\theta|x)$ 最小的行动 $a = a(x)$. 这一原则常称为**后验风险最小的原则**, 而 $\int_\Theta L(\theta, a)P(d\theta|x)$ 可视为采取行动 a 导致的后验风险. 这样得到的估计一般是非随机化的. 因此我们今后将只需考虑非随机化的 Bayes 估计. 在一般情形涉及的解的存在性和可测性问题, 已超出本书的范围, 读者可参考陈希孺 (1981, 1997) 的第 145–150 页.

2.3.2 一些重要情形的 Bayes 估计

下面我们举例说明如何求解 Bayes 估计问题.

例 2.11 设样本空间是欧氏的. 采用加权的平方损失 $L(\theta, a) = \lambda(\theta)\big(g(\theta) - a\big)^2$, 此处 $\lambda(\cdot) > 0$ 和 $g(\cdot)$ 都是已知的 $\mathcal{B}_\Theta$-可测函数. 损失函数的形式表明问题是估计 $g(\theta)$. 设 ξ 为先验分布, 样本为 $X = x$. 令

$$K(a, x) = \int_\Theta \lambda(\theta)\big(g(\theta) - a\big)^2 P(d\theta|x).$$

不难证明, 若在 $\mathcal{A}$ 中存在两点 $a_1 \neq a_2$, 使 $K(a_1, x)$ 和 $K(a_2, x)$ 均为有限, 则 $K(\cdot, x)$ 在 $\mathcal{A}$ 中处处有限, 且 $\mathrm{E}[\lambda(\theta)|g(\theta)|^i|x] < \infty, i = 0, 1, 2$. 根据后验风险最小的原则, 并区别仅有的以下三种情况, 可求得 Bayes 估计 $\delta_\xi(X)$:

(i) 对 $\forall a \in \mathcal{A}$, 都有 $K(a,x)=\infty$, 此时可取 $\delta_\xi(x)=a_0$, 其中 a_0 为 $\mathcal{A}$ 中选定的任一行动;

(ii) $K(a,x)$ 只在一点 $a=a_x$ 处为有限, 此时 $\delta_\xi(x)=a_x$;

(iii) 对 $\forall a \in \mathcal{A}$, 都有 $K(a,x)$ 有限, 此时 $\delta_\xi(x)=\mathrm{E}[\lambda(\theta)g(\theta)|x]/\mathrm{E}[\lambda(\theta)|x]$.

特别地, 若损失函数取为通常的平方损失 $(\theta-a)^2$, 则 θ 的 Bayes 估计即为其后验均值, 确切地说, Bayes 估计为

$$\delta_\xi(x)=\begin{cases}\mathcal{A}\text{ 中任意选定的 }a_0, & \text{若 } \mathrm{E}(\theta^2|x)=\infty,\\ \mathrm{E}(\theta|x)=\displaystyle\int_\Theta \theta P(d\theta|x), & \text{若 } \mathrm{E}(\theta^2|x)<\infty.\end{cases} \tag{2.40}$$

为了下面的需要, 我们引进下述定义.

称 μ 为一个一维随机变量Z 或其分布 F 的一个**中位数** (median), 若 $P(Z\leqslant\mu)\geqslant 1/2$, 且 $P(Z\geqslant\mu)\geqslant 1/2$, 即 $F(\mu-)\leqslant 1/2\leqslant F(\mu)$. 一个分布的中位数通常不唯一, 其全体形成一个有界闭区间. 不难证明, μ 为 Z 的一个中位数, 当且仅当 $\mathrm{E}(|Z-t|-|Z|)$ 在 $t=\mu$ 处取得最小值. 类似地, 给定 $0<p<1$, 称 μ_p 为 Z 或其分布 F 的一个 p 分位数, 若 $F(\mu_p-)\leqslant p\leqslant F(\mu_p)$. 定义 $\rho_p(z)=2z[p-I(z<0)]$, 其中 $I(B)$ 为集合 B 的示性函数. 则 μ_p 为 Z 的一个 p **分位数** (quantile), 当且仅当 $\mathrm{E}[\rho_p(Z-t)-\rho_p(Z)]$ 在 $t=\mu_p$ 处取得最小值. 我们要特别指出, $\rho_p(\cdot)$ 为一凸函数, 且当 $p=1/2$ 时即为绝对值函数.

例 2.12 在例 2.11 中仍取 ξ 为先验分布, 而将损失函数改为 $L(\theta,a)=|\theta-a|$, 由此可以看出问题是估计 θ. 记 $K(a,x)=\mathrm{E}(|\theta-a|\ |x)$. 由后验风险最小的原则, 考虑到上述说明, 可得 θ 的 Bayes 估计即为后验分布的中位数, 确切地说, Bayes 估计为

$$\delta_\xi(x)=\begin{cases}\mathcal{A}\text{ 中任意选定的 }a_0, & \text{若 } \mathrm{E}(|\theta|\ |x)=\infty,\\ \text{后验分布 }P(\cdot|x)\text{ 的中位数}, & \text{若 } \mathrm{E}(|\theta|\ |x)<\infty.\end{cases} \tag{2.41}$$

类似地, 我们也可以考虑损失函数为 $\lambda(\theta)|g(\theta)-a|$ 或者 $\lambda(\theta)\rho_p\big(g(\theta)-a\big)$ 的一般情形.

例 2.13 设 X 服从二项分布 $B(n,\theta)$, $0\leqslant\theta\leqslant 1$, 今有样本 x, 如何估计 θ? 通常我们用 x/n 来估计它. 现在我们采用 Bayes 方法. 损失函数取为 $(\theta-a)^2$. 即使对 θ 一无所知, 我们也可以根据所谓 "同等无知" 的原则, 认为 θ 以同等机会取 $(0,1)$ 区间内各种值, 因而选取先验分布为 $R(0,1)$. 此时 X 的边缘分布为

$$P^X(x)=\int_0^1\binom{n}{x}\theta^x(1-\theta)^{n-x}d\theta=\frac{1}{n+1},\quad x=0,1,\cdots,n,$$

此即 $\{0,1,\cdots,n\}$ 上的均匀分布, 而 θ 的后验分布为

$$P(d\theta|x)=\binom{n}{x}\theta^x(1-\theta)^{n-x}d\theta\Big/P^X(x)=(n+1)\binom{n}{x}\theta^x(1-\theta)^{n-x}d\theta.$$

这个分布即为 $Be(x+1,n-x+1)$. 由例 2.11, θ 的 Bayes 估计可取为后验均值, 即 $\delta_\xi(x)=(x+1)/(n+2)$. 将 p 看作 Bernoulli 试验中某事件出现的概率, n 为试验重

复的次数. 设想有两种情况, 其一为该事件在一百次试验中未出现, 其二为仅在两次试验中未出现, 则前一情形将 θ 估低些而后一情形估高些较为合理. 在现在的情形, 当 $x=0$ 时, $\delta_\xi(x)$ 用 $1/(n+2)$ 估计 θ, 与此要求相符; 而用 x/n 估计时, 估计值 0 与 n 无关, 显得不大自然. 当 $x=1$ 时也有类似的情况.

上面选取的先验分布就是一种无信息先验 (noninformative prior) 分布. 关于无信息先验, 我们将在 2.3.7 节作简要介绍. 详细论述可参看 Berger (1985) 的 3.3 节和 6.6 节.

现将先验分布 ξ 改为一般的 $Be(a,b), a\geqslant 0, b\geqslant 0$, 而保持上面的其他假定不变. 联合分布 $P(X=x,d\theta)$ 可以表示为

$$\binom{n}{x}\theta^x(1-\theta)^{n-x}\cdot\theta^{a-1}(1-\theta)^{b-1}d\theta\Big/\beta(a,b),\ x=0,\cdots,n, 0<\theta<1.$$

因而在得到样本 $X=x$ 时, θ 的后验分布为 $P(d\theta|x)=c\theta^{x+a-1}(1-\theta)^{n-x+b-1}d\theta$, 此分布即为 $Be(x+a,n-x+b)$ 且 $c=\Gamma(n+a+b)/[\Gamma(x+a)\Gamma(n-x+b)]$. θ 的 Bayes 估计即为其后验均值

$$\delta_\xi(x)=(x+a)/(n+a+b), \tag{2.42}$$

其风险函数为

$$R(\theta,\delta_\xi)=\mathrm{E}[(\delta_\xi(X)-\theta)^2|\theta]=\frac{n\theta(1-\theta)+[(a+b)\theta-a]^2}{(n+a+b)^2},$$

Bayes 风险为

$$R_\xi(\delta_\xi)=ab/[(n+a+b)(a+b)(a+b+1)].$$

而当 $a=b=\sqrt{n}/2$ 时, Bayes 估计 的风险函数

$$R(\theta,\delta_\xi)=n/[4(n+\sqrt{n})^2] \tag{2.43}$$

为一常数.

又若取绝对值损失 $|\theta-a|$, 则 Bayes 估计为后验分布 $Be(x+a,n-x+b)$ 的中位数, 一般需要查表才能确定.

例 2.14 设 $X_1,\cdots,X_n$ 为从正态总体 $N(\theta,\sigma^2)$ 中抽取的 i.i.d. 样本, 其中 $\sigma^2>0$ 已知, $-\infty<\theta<\infty$ 为未知参数, 要估计 θ. 损失函数为 $(\theta-a)^2$, 先验分布 ξ 为 $N(\nu,\gamma^2)$, $-\infty<\nu<\infty$, $\gamma^2>0$. 因 $\bar{X}$ 为完全充分统计量, 可以只考虑基于它的估计, 其抽样分布为 $N(\theta,\sigma^2/n)$. 设其样本值为 $\bar{x}$. 由 (2.34), 知后验分布为 $N(\theta_x,\sigma_x^2)$, 其中

$$\theta_x=(n\sigma^{-2}\bar{x}+\gamma^{-2}\nu)/(n\sigma^{-2}+\gamma^{-2}),\quad \sigma_x^2=1/(n\sigma^{-2}+\gamma^{-2}).$$

θ 的 Bayes 估计, 即其后验均值为

$$\delta_\xi(x)=\theta_x=\frac{n\sigma^{-2}}{n\sigma^{-2}+\gamma^{-2}}\bar{x}+\frac{\gamma^{-2}}{n\sigma^{-2}+\gamma^{-2}}\nu. \tag{2.44}$$

不难算得其风险函数为

$$R(\theta,\delta_\xi)=\frac{n\sigma^{-2}+\gamma^{-4}(\nu-\theta)^2}{(n\sigma^{-2}+\gamma^{-2})^2}.$$

这一估计形象地反映了一个基本事实, 即后验分布综合了先验信息 (ξ) 和样本信息. 若只考虑样本信息, 则 θ 的一个合理的估计为 $\bar{x}$. 如果没有样本而只利用先验分布 $N(\nu,\gamma^2)$ 提供的信息, 则其均值 ν 可取为 θ 的一个估计. δ_ξ 则是两者的一种加权平均, 权的大小取决于两种信息重要性的对比: 样本量 n 愈大, σ^2 愈小, 则样本信息愈大; γ^2 愈小, 则先验信息愈大.

现设我们取先验分布 ξ_k 为 $N(0,\gamma_k^2)$, 相应的 Bayes 估计为 $\delta_{\xi_k}(X)$. 如果当 $k\to\infty$ 时 $\gamma_k\to\infty$, 则容易看出, $\delta_{\xi_k}(X)\to\bar{X}$ a.s., 且 δ_{ξ_k} 的 Bayes 风险 $R_{\xi_k}(\delta_{\xi_k})$ 趋于 $\bar{X}$ 的常数风险 σ^2/n , 这一事实在后面研究 minimax 估计时有用.

例 2.15 设 X 表示考察的元件的寿命, $X\sim\{E(0,1/\theta),\theta>0\}$, 即 X 的密度函数为 $f(x|\theta)=\theta\mathrm{e}^{-x\theta}I(x>0)$, θ 为元件的失效率. 设 $X_1,\cdots,X_n$ 为取自该总体的 i.i.d. 样本, 要估计元件在指定时刻 t 之后失效的概率 $P_\theta(X>t)=\mathrm{e}^{-t\theta}$. 取平方损失 $(\mathrm{e}^{-t\theta}-a)^2$. θ 的先验分布 ξ 取为 Gamma 分布 $G(\kappa,1/\lambda),\kappa>0,\lambda>0$. 其密度函数为

$$g(\theta;\kappa,1/\lambda)=\frac{\lambda^\kappa}{\Gamma(\kappa)}\theta^{\kappa-1}\mathrm{e}^{-\lambda\theta}I(\theta>0).$$

记 $t_n=\sum_{i=1}^n x_i$, 则 $(X_1,\cdots,X_n,\theta)$ 的联合密度为

$$c_1\theta^n\exp\{-t_n\theta\}\theta^{\kappa-1}\exp\{-\lambda\theta\}I(x_1>0,\cdots,x_n>0,\theta>0),$$

其中 c_1 为一常数. 容易看出, 在取得样本 $x_1,\cdots,x_n$ 后 θ 的后验密度为

$$c(x)\theta^{n+\kappa-1}\exp\{-(t_n+\lambda)\theta\}I(\theta>0),$$

其中 $c(x)=(t_n+\lambda)^{n+\kappa}/\Gamma(n+\kappa)$, 因而 θ 的后验分布为 $G\big(n+\kappa,(t_n+\lambda)^{-1}\big)$. $\mathrm{e}^{-t\theta}$ 的 Bayes 估计, 即其后验均值为

$$\begin{aligned}&\frac{(t_n+\lambda)^{n+\kappa}}{\Gamma(n+\kappa)}\int_0^\infty\theta^{n+\kappa-1}\exp\{-(t+t_n+\lambda)\theta\}d\theta\\&=\frac{(t_n+\lambda)^{n+\kappa}}{(t+t_n+\lambda)^{n+\kappa}}=\Big[1+\frac{t}{t_n+\lambda}\Big]^{-(n+\kappa)}.\end{aligned}$$

2.3.3 共轭先验分布族

在例 2.13 — 例 2.15 中, 有关的先验分布族都具有以下性质: 对任何样本 x, 后验分布 $P(\cdot|x)$ 仍在此族中. 这导致了下述定义.

定义 2.10 设有样本空间 $(\mathcal{X},\mathcal{B},\mathcal{P}),\mathcal{P}=\{P_\theta,\theta\in\Theta\}$. 称 $\mathcal{B}_\Theta$ 上的分布族 $\mathcal{G}$ 是样本分布族 $\mathcal{P}$ 的一个**共轭先验分布族** (conjugate family of prior distribution), 如果当 θ 的先验分布 $\xi\in\mathcal{G}$ 时, 对任何样本 $X=x$ (或除去一例外的 $\mathcal{P}$-零概集), 后验分布 $P(\cdot|x)$ 也属于 $\mathcal{G}$.

据此定义, 在二项样本分布族的情形, $\{Be(a,b),a\geqslant 0,b\geqslant 0\}$ 即为一个共轭先验分布族. 在有关问题中, 如果存在一个足够大的共轭先验分布族, 我们可从中选取一个适当的分布作为先验分布, 这不仅是数学上的一种简化, 而且不失为一种合适的建模方法.

为了叙述方便, 我们特作如下约定: 设 $\mathcal{P}$ 为一概率测度族, $Q(x)$ 是关于 x 的一个命题, 称 $Q(x)$ a.e. $\mathcal{P}$ 成立, 如果对于每个 $P\in\mathcal{P}$ 都有 $Q(x)$, a.e. P.

今设 $dP_\theta(x)=f(x,\theta)d\mu(x)$, $x\in\mathcal{X},\theta\in\Theta$, 其中 μ 为 $\mathcal{B}$ 上的一个 σ-有限测度, f 为 $\mathcal{B}\times\mathcal{B}_\Theta$-可测函数. 设 $X_1,\cdots,X_n$ 为从上述分布族中取出的 i.i.d. 样本. 记 $P_\theta^n=\prod_{i=1}^n P_\theta$, $\mathcal{P}^n=\{P_\theta^n,\ \theta\in\Theta\}, X^n=(X_1,\cdots,X_n)^\tau$. 类似地, 可定义 μ^n, $\mathcal{X}^n$, $\mathcal{B}^n$等.

假设对每个 n, 存在充分统计量 $T_n=T_n(X^n)$. 依因子分解定理, 有

$$\prod_{i=1}^n f(x_i,\theta)=g_n[T_n(x^n),\theta]h_n(x^n),\ n=1,2,\cdots.$$

定理 2.15 在上述架构下, 假定对 $\forall x^n\in\mathcal{X}^n$ (可以有一个例外的 $\mathcal{P}^n$-零概集, 下同), 有

$$0<\int_\Theta g_n\big(T_n(x^n),\theta\big)d\nu(\theta)<\infty,$$

此处 ν 为 $\mathcal{B}_\Theta$ 上给定的一个 σ-有限测度, 则 $\mathcal{B}_\Theta$ 上的分布族

$$\mathcal{G}_\nu=\Big\{\frac{g_n[T_n(x^n),\theta]d\nu(\theta)}{\int_\Theta g_n[T_n(x^n),\psi]d\nu(\psi)},\ x^n\in\mathcal{X}^n,\ n=1,2,\cdots\Big\}$$

构成一个共轭先验分布族.

证明 在 $\mathcal{G}_\nu$ 族中取定一个分布

$$c(\theta)d\nu(\theta)=cg_m[T_m(z^m),\theta]d\nu(\theta)$$

作为 θ 的先验分布, 且设有了样本 $X^n=x^n$, 则由 (2.34), 后验分布为

$$\begin{aligned}q_n(x^n,\theta)d\nu(\theta)&=\frac{g_m[T_m(z^m),\theta]g_n[T_n(x^n),\theta]d\nu(\theta)}{\int_\Theta g_m[T_m(z^m),\psi]g_n[T_n(x^n),\psi]d\nu(\psi)}\\&=\frac{g_{m+n}[T_{m+n}(z^m,x^n),\theta]d\nu(\theta)}{\int_\Theta g_{m+n}[T_{m+n}(z^m,x^n),\psi]d\nu(\psi)},\end{aligned}$$

为得到最后一个等式, 可对 $\prod_{i=1}^m f(z_i,\theta)$ 和 $\prod_{j=1}^n f(x_j,\theta)$ 的乘积应用因子分解定理, 并对两者分别应用之. 这就是所要证明的. ■

在实际应用中, 有时可将由该定理确定的共轭分布族适当扩大, 使其仍不失共轭性. 上面的三个例子中的共轭先验分布族都可通过这一途径推演出来. 除了下面的例子, 我们还将举出其他一些例子放在本章后面的习题中.

例 2.16 设 $X_1,X_2,\cdots$, 为 i.i.d. 正态 $N(0,\sigma^2)$ 序列, $\sigma>0$. 为寻求共轭先验分布族, 取 $d\nu(\sigma)=d\sigma$. 样本 $(X_1,\cdots,X_n)$ 的分布为 $(\sqrt{2\pi}\sigma)^{-n}\exp(-t_n/2\sigma^2)dx_1\cdots dx_n$, 此处, $t_n:=\sum_{i=1}^n x_i^2$, $T_n:=\sum_{i=1}^n X_i^2$ 为充分统计量. 因 $T_n>0$ 以概率 1 成立, 故为了使定理 2.15 中的条件

$$\int_0^\infty \sigma^{-n}\exp\{-t_n/(2\sigma^2)\}d\sigma<\infty$$

成立, 只需 $n>1$. 由定理 2.15 可得共轭先验分布族

$$\{c(n,\alpha)\sigma^{-n}\exp\{-\alpha/(2\sigma^2)\}I(\sigma>0)d\sigma:\alpha>0,n=2,3,\cdots\},\tag{2.45}$$

$c(n,\alpha)$ 为一仅与 n,α 有关的常数. 此分布族可扩大为以下的分布族而保持其共轭性:

$$\{c(\kappa,\alpha)\sigma^{-(2\kappa+1)}\exp\{-\alpha/(2\sigma^2)\}I(\sigma>0)d\sigma:\kappa>0,\alpha>0\},\tag{2.46}$$

此处, $c(\kappa,\alpha)$ 为一仅与 κ,α 有关的常数.

若记 $1/\omega=2\sigma^2$, 则当 $\sigma\sim$ (2.46) 时 $\omega\sim G(\kappa,1/\alpha),\kappa>0,\alpha>0$.

今设 $X^n=(X_1,\cdots,X_n)$ 为样本, 要在平方损失 $(1/\omega-a)^2$ 之下估计 $1/\omega=2\sigma^2$. ω 的先验分布 ξ 取为上述 $G(\kappa,1/\alpha)$, 即 $\omega\sim[\alpha^\kappa/\Gamma(\kappa)]\omega^{\kappa-1}\exp(-\alpha\omega)d\omega$, 则后验分布为 $\omega|x^n\sim G(n/2+\kappa,1/(t_n+\alpha))$. 注意到若 $\omega\sim G(\kappa,1/\alpha)$, 则 $\mathrm{E}(\omega)=\kappa/\alpha$, $\mathrm{E}(\omega^2)=\kappa(\kappa+1)/\alpha^2$, $\mathrm{E}(1/\omega)=\alpha/(\kappa-1)$, 故 $1/\omega=2\sigma^2$ 的 Bayes 估计为

$$\delta_\xi(x^n)=\mathrm{E}(1/\omega|x^n)=(t_n+\alpha)/(n/2+\kappa-1),$$

其中 $t_n=\sum_{i=1}^n x_i^2$. 需要注意的是, 为使 Bayes 估计的 Bayes 风险为有限, 必须而且只需 $\mathrm{E}(1/\omega^2)<\infty$, 即 $\kappa>2$.

我们更感兴趣的是, 将损失函数取为在刻度变换下不变的, $L(\omega,a)=(a\omega-1)^2=[a/(2\sigma^2)-1]^2$. 此时不难推得 $2\sigma^2$ 的 Bayes 估计为

$$\delta_\xi(x^n)=\mathrm{E}(\omega|x^n)/\mathrm{E}(\omega^2|x^n)=(t_n+\alpha)/(n/2+\kappa+1).$$

例 2.17 设 $X_1,\cdots,X_n$ 为取自正态总体 $N(\theta,\sigma^2)$ 的 i.i.d. 样本, 均值方差均未知. 要寻求共轭先验分布族. 易见样本联合分布可表示为

$$(\sqrt{2\pi}\sigma)^{-n}\exp\{-(t_n+n(\bar{x}_n-\theta)^2)/2\sigma^2\}dx_1\cdots dx_n,$$

此处 $\bar{x}_n=\sum_{i=1}^n x_i/n$, $t_n=\sum_{i=1}^n(x_i-\bar{x}_n)^2$, 而 $(\sum_{i=1}^n X_i/n,\sum_{i=1}^n(X_i-\bar{X}_n)^2)$ 为充分统计量. 于是得到 (θ,σ) 的一个共轭先验分布族: $\sigma\sim$ (2.45), 而在给定 σ 的条件下, θ 的条件分布为 $N(\mu,\sigma^2/n)$. 该共轭先验分布族可以扩大如下: $\omega:=1/(2\sigma^2)\sim G(\kappa,1/\alpha),\kappa>0$, $\alpha>0$; 而在给定 σ 时, θ 的条件分布为 $N(\mu,\sigma^2/\gamma)$, $\mu\in(-\infty,\infty),\gamma>0$.

事实上通过直接计算可以证明, 若取这样一个分布作为 (θ,σ) 的先验分布, 则其后验分布为: $\omega=1/(2\sigma^2)\sim G\big(\kappa+n/2,1/[\alpha+t_n+n\gamma(\bar{x}_n-\mu)^2/(n+\gamma)]\big)$, 而在给定 σ 时, θ 的条件分布为 $N\big((n\bar{x}_n+\gamma\mu)/(n+\gamma),\sigma^2/(n+\gamma)\big)$. 此处我们将这一计算留给读者作为练习.

2.3.4 广义 Bayes 估计

无论是从理论还是从应用的角度考虑, Bayes 估计类都需要适当地扩大, 因为一些常见的估计不是 Bayes 估计.

例 2.18 设 $X_1,\cdots,X_n$ 为取自正态总体 $N(\theta,\sigma^2)$ 的 i.i.d. 样本, $\sigma^2>0$ 已知, $-\infty<\theta<\infty$, 损失函数为 $L(\theta,a)=(\theta-a)^2$. 今证不论 θ 取何种先验分布 ξ, 常见的估计 $\bar{X}=\sum_{i=1}^n X_i/n$ 绝非 Bayes 估计.

不失一般性, 可设 $n=1$ 且取 $\sigma^2=1$. 简记 $X_1=X$. 若 X 为 Bayes 估计, 则由 (2.40), 应有

$$x=\frac{\int_{-\infty}^{\infty}\theta\exp\{-(x-\theta)^2/2\}\,d\xi(\theta)}{\int_{-\infty}^{\infty}\exp\{-(x-\theta)^2/2\}\,d\xi(\theta)},\ \text{a.e. } L \tag{2.47}$$

成立, 此处 L 为 Lebesgue 测度. 记

$$h(x)=\int_{-\infty}^{\infty}\exp\{-(x-\theta)^2/2\}d\xi(\theta), \tag{2.48}$$

依指数族的性质, $h'(x)$ 连续且可通过在积分号下求导来计算, 故

$$h'(x)=\int_{-\infty}^{\infty}(\theta-x)\exp\{-(x-\theta)^2/2\}d\xi(\theta).$$

将 (2.47) 代入上式, 即得 $h'(x)=0,\text{a.e. } L$. 由函数 h' 的连续性知 $h'(x)\equiv 0$ 从而 $h(x)\equiv$ 常数. 对 (2.48) 用控制收敛定理, 得 $\lim_{x\to\infty}h(x)=0$, 于是 $h(x)\equiv 0$. 但 (2.48) 中被积函数恒为正且 ξ 为概率测度, 故这是不可能的. 这个矛盾即证明了 X 确非 Bayes 估计.

1. 决策函数列的正则收敛 —— 广义 Bayes 估计定义之一

由例 2.14 可知, 尽管 $\bar{X}$ 并非 Bayes 估计, 但当取 $\gamma:=\gamma_n\to\infty$ 时, 它却是一串 Bayes 估计的极限. 这启示我们可在某种极限意义下对 Bayes 估计作一自然的推广. 为此, 设样本空间为 $(\mathcal{X},\mathcal{B})$, 分布族为 $\{P_\theta,\theta\in\Theta\}$, 且设行动空间 $(\mathcal{A},\mathcal{B}_\mathcal{A})$ 为欧氏的.

定义 2.11 (决策函数列的正则收敛) 设 $\delta,\delta_1,\delta_2,\cdots$ 为一串非随机化决策函数, 若对 $\mathcal{A}$ 上任何有界连续函数 $u(\cdot)$, 以及 $\mathcal{X}$ 上任何满足条件

$$\int_\mathcal{X}|g(x)|dP_\theta(x)<\infty,\ \text{对 }\forall\ \theta\in\Theta$$

的函数 g, 必有

$$\lim_{n\to\infty}\int_\mathcal{X}u(\delta_n(x))g(x)dP_\theta(x)=\int_\mathcal{X}u(\delta(x))g(x)dP_\theta(x),\quad\forall\ \theta\in\Theta,$$

则称 $\{\delta_n\}$ 正则收敛于 δ, 并记为 $\delta_n\xrightarrow{r}\delta$.

在随机化决策函数的情形, 正则收敛的定义为

$$\lim_{n\to\infty}\int_\mathcal{X}\int_\mathcal{A}u(a)\delta_n(da|x)g(x)dP_\theta(x)=\int_\mathcal{X}\int_\mathcal{A}u(a)\delta(da|x)g(x)dP_\theta(x),\ \forall\ \theta\in\Theta.$$

不难看出, 这包含了非随机化情形作为一个特例. 由控制收敛定理, 在非随机化情形, 若对任何 P_θ, 有 $\delta_n(X)\xrightarrow{\text{pr.}}\delta(X)$, 则 $\delta_n\xrightarrow{r}\delta$.

定义 2.12 在一定的样本空间、分布族、行动空间、损失函数之下, 设对任何 n, 决策函数 δ_n 为关于先验分布 ξ_n 的 Bayes 解, 且 $\delta_n\xrightarrow{r}\delta$. 若 δ 自身不是 Bayes 解, 则称它为一个**广义 Bayes 解**.

关于广义 Bayes 解的一个主要结果, 是在很宽广的条件下, Wald (1950) 证明了, 所有 Bayes 解和广义 Bayes 解全体构成了统计决策函数的一个完全类. 对 Bayes 决

策函数类添加缺漏的重要成员, 使之成为一个完全类或本质完全类, 是 Wald 这本经典著作讨论的一个基本问题, 多年来备受一些理论学者的关注. 下面我们不加证明地引述其中的一个结果.

定理 2.16 设一统计决策问题中的参数空间和行动空间都是欧氏空间中的有界闭集, 样本分布为 $f(x,\theta)d\mu$, 其中 μ 为 Lebesgue 测度或计数测度. 对每个给定的 θ, 损失函数 $L(\theta,\cdot)$ 为行动空间上的连续函数, 且任一决策函数的风险函数为参数空间上的连续函数, 则所有 Bayes 决策函数的类为一完全类.

参看 Berger (1985) 第 546 页的定理 12, 及该处所引文献.

2. 广义先验分布 —— 广义 Bayes 估计定义之二

仍设 X 的样本空间为 $(\mathcal{X},\mathcal{B})$, 分布族为 $\{P_\theta,\theta\in\Theta\}$, 参数空间和行动空间分别为 $(\Theta,\mathcal{B}_\Theta)$ 和 $(\mathcal{A},\mathcal{B}_\mathcal{A})$. 假定 $dP_\theta(x)=f(x,\theta)d\mu(x),\theta\in\Theta$, 此处, μ 为样本空间上的一个 σ-有限测度. 如前所述, 从形式上讲, 若将 ξ 取为 $\mathcal{B}_\Theta$ 上的任一测度, 仍可按定义 2.7 定义 Bayes 解. 问题在于, 对先验分布作了这种推广之后, "后验风险最小"的原则是否仍能保持? 事实上, 由 (2.34) 定义的 $P(\cdot|x)$ 仍构成一个概率分布, 只要

$$0<\int_\Theta f(x,\theta)d\xi(\theta)<\infty, \tag{2.49}$$

而不需要假定 $\xi(\Theta)=1$. 这些考虑导致下述定义.

定义 2.13 沿用上述符号和假定. 设 ξ 为 $\mathcal{B}_\Theta$ 上的一个测度, $\xi(\Theta)=\infty$, 且存在集合 $N\in\mathcal{B}$, $\int_\Theta P_\theta(N)d\xi(\theta)=0$, 使 (2.49) 对所有 $x\in N^c$ 成立, 则称 ξ 为一**广义先验分布** (Generalized prior distribution).

对于 $0<\xi(\Theta)=C<\infty$ 的情形, 当用 ξ/C 代替 ξ 时, 就回到通常先验分布的情形, 因而定义中只需考虑 $\xi(\Theta)=\infty$ 的情形.

下面的定义与定义 2.7 在形式上毫无区别.

定义 2.14 沿用上述符号和假定. 设 ξ 为 θ 的一个广义先验分布, 则由

$$R_\xi(\delta)=\int_\Theta R(\theta,\delta)d\xi(\theta)$$

定义的$R_\xi(\delta)$ 称为决策函数 δ 关于广义先验分布 ξ 的 **Bayes 风险**. 如果某个决策函数 δ_ξ 满足

$$R_\xi(\delta_\xi)\leqslant R_\xi(\delta),\ \text{对所有 } \delta,$$

则称 δ_ξ 为**相应于广义先验分布 ξ 的广义 Bayes 解**.

在广义先验分布的场合, 后验分布已不能当作条件分布来定义, 而只能仿照 (2.34) 形式地定义:

$$P(d\theta|x)=f(x,\theta)d\xi(\theta)\Big/\int_\Theta f(x,\theta)d\xi(\theta).$$

在 (2.49) 成立的条件下, 它确实是一个概率分布. 此时, 决策函数 δ 的后验风险仍可按定义 2.9 中的 (2.37) 定义:

$$R_\xi(\delta,x)=\begin{cases}\displaystyle\int_\Theta L(\theta,\delta(x))P(d\theta|x), & (\delta \text{ 为非随机化})\\ \displaystyle\int_{\mathcal{A}}\delta(x,da)\int_\Theta L(\theta,a)P(d\theta|x). & (\delta \text{ 为随机化})\end{cases}$$

下面的定理表明, 在广义先验分布下, "后验风险最小" 的原则仍然适用.

定理 2.17 *设存在集合 $N\in\mathcal{B},\int_\Theta P_\theta(N)d\xi(\theta)=0$, 当 $x\notin N$ 时,*

$$\int_\Theta L(\theta,\delta(x))P(d\theta|x)=\inf_{a\in\mathcal{A}}\int_\Theta L(\theta,a)P(d\theta|x), \tag{2.50}$$

则 δ 就是在广义先验分布 ξ 之下的广义 Bayes 解.

定理证明留给读者作为练习.

以上提出了广义 Bayes 解的两个定义. 正如例 2.18 和例 2.14 所示, 在正态样本和平方损失之下, $\bar{X}$ 不是样本均值 θ 的一个 Bayes 估计, 但可取作 θ 的一串 Bayes 估计在正则收敛意义下的极限. 另一方面, $\bar{X}$ 亦为对应于广义先验分布 $d\xi(\theta)=d\theta$ 的广义 Bayes 解, 即在此情形之下, $\bar{X}$ 同时满足广义 Bayes 解的两个定义. 至于一般情形两种广义 Bayes 解的关系, 即在何种条件下两者一致的问题, 有些学者研究过, 具体内容很复杂, 已超出本书范围, 有兴趣的读者可参看 Sacks (1963) 和 Farrell (1966).

2.3.5 经验 Bayes 估计

经验 Bayes (Empirical Bayes) 方法, 简称 EB 方法, 是 Robbins 在 1955 年提出的, 其目的是为了解决使用 Bayes 方法时, 因不知道先验分布而产生的困难. 粗略地说, EB 方法就是前面提到过的, 通过历史资料 (即 "经验") 决定先验分布的方法. 这个方法使用的前提是, 参数 θ 确是通常意义下的随机变量, 其分布 ξ 在历史资料积累的过程中保持不变. 本章开头提到的那个估计废品率的问题, 即为适合这一条件的典型例子.

下面我们来说明 EB 方法的基本思想和严格定义. 为此, 设有了决策问题的诸要素 $(\mathcal{X},\mathcal{B},P_\theta),\theta\in\Theta,(\Theta,\mathcal{B}_\Theta),(\mathcal{A},\mathcal{B}_\mathcal{A}),L(\theta,a)$ 和先验分布 ξ, 此处 $X|\theta\sim P_\theta,\theta\sim\xi$. 分别用 $\delta_\xi(X)$ 和 $R_\xi(\delta_\xi)$ 表示在先验分布 ξ 之下的 Bayes 解及其 Bayes 风险. 通常只知道 ξ 属于某个分布族 Ξ.

假定这个决策问题多次出现, 每次出现时, 上述诸要素不变. 这句话的含义是, 若以 $(X_1,\theta_1),\cdots,(X_n,\theta_n)$ 记 (X,θ) 在历史上的重复, 则它们与当前的 (X,θ) 独立同分布. 但在此处, $\theta,\theta_1,\cdots,\theta_n$ 是不可观测的, 只有 $X,X_1,\cdots,X_n$ 是可观测的, 它们独立同分布, 其分布与先验分布 ξ 有关, 记为 P_ξ^X, $P_\xi^X(\cdot)=\int_\Theta P_\theta(\cdot)d\xi(\theta)$. X 称为当前样本, $X_1,\cdots,X_n$ 称为历史样本, 简记为 X^n. 样本也常用小写字母来表示.

在上述架构下, 我们关心的问题是, 如何利用历史样本 X^n 对与当前样本 X 相匹配的 θ 作出判决. 此时, 相应的决策函数记为 $\delta_n(X;X^n):=\delta_n(X;X_1,\cdots,X_n)$, 称

为 EB 决策函数, 当用于估计问题时就称为 EB 估计, 其全面 Bayes 风险定义为

$$R_\xi(\delta_n) = \mathrm{E}\big[L\big(\theta, \delta_n(X; X^n)\big)\big] := \mathrm{E}[R_\xi(\delta_n|X^n)], \tag{2.51}$$

其中条件 Bayes 风险

$$R_\xi(\delta_n|x^n) = \mathrm{E}\big[L\big(\theta, \delta_n(X; X^n)\big)|x^n\big]. \tag{2.52}$$

对固定的 x^n , $\delta_n(X; x^n)$ 作为一个决策函数, 其 Bayes 风险 $R_\xi(\delta_n|x^n)$ 不小于 Bayes 解的 Bayes 风险. 故对 δ_n 的全面 Bayes 风险, 有

$$R_\xi(\delta_n) \geqslant R_\xi(\delta_\xi). \tag{2.53}$$

我们自然希望经验 Bayes 决策函数的全面 Bayes 风险接近 $R_\xi(\delta_\xi)$. 为此, 我们引进如下的定义.

定义 2.15 任何同时依赖于历史样本 X^n 和当前样本 X 的决策函数 $\delta_n(X; X^n)$ 都称为经验 Bayes 决策函数. 设先验分布 ξ 属于分布族 Ξ. 经验 Bayes 决策函数 δ_n 称为关于先验分布族 Ξ 是渐近最优的(简记为 a.o.), 如果对任何 $\xi \in \Xi$, 都有

$$\lim_{n\to\infty} R_\xi(\delta_n) = R_\xi(\delta_\xi). \tag{2.54}$$

在一般情况下, 当构造经验 Bayes 决策函数时, 历史样本与当前样本的作用是不一样的. 历史样本用于获得关于先验分布 ξ 的信息以帮助选定一个合适的决策函数 $\delta_n(\cdot; X^n)$, 而当前样本 X 则主要用于对当前参数值 θ 作出推断. 下面举例加以说明.

例 2.19 设 $X \sim N(\theta, 1)$, 损失函数 $L(\theta, a) = (\theta - a)^2$. θ 的先验分布 $\xi \in \Xi = \{N(0, \gamma^2),\ \gamma > 0\}$. 由例 2.14 , θ 的 Bayes 估计为

$$\delta_\xi(X) = \big[\gamma^2/(1+\gamma^2)\big]X,$$

其 Bayes 风险为 $\gamma^2/(1+\gamma^2)$.

设 $X_1, \cdots, X_n$ 和 X 分别为历史样本和当前样本, 要构造 θ 的经验 Bayes 估计. 容易验证, X 的边缘分布 P_ξ^X 为 $N(0, 1+\gamma^2)$. 故可得 γ^2 的一个估计

$$\hat{\gamma}_n^2 = \frac{1}{n}\sum_{i=1}^n X_i^2 - 1.$$

如前所述, 在先验分布 $N(0, \hat{\gamma}_n^2)$ 之下, 基于样本 X 得到的 Bayes 估计为

$$\delta_n(X; X^n) = \big[\hat{\gamma}_n^2/(1+\hat{\gamma}_n^2)\big]X,$$

它在 X^n 给定时的条件 Bayes 风险为 $\hat{\gamma}_n^2/(1+\hat{\gamma}_n^2)$, 故其全面 Bayes 风险为

$$R_\xi(\delta_n) = \mathrm{E}\big[\hat{\gamma}_n^2/(1+\hat{\gamma}_n^2)\big].$$

由大数定律, 以概率为 1 地成立

$$\hat{\gamma}_n^2 \to (1+\gamma^2) - 1 = \gamma^2.$$

由有界收敛定理即得

$$\lim_{n\to\infty} R_\xi(\delta_n) = \gamma^2/(1+\gamma^2),$$

因而 δ_n 即为关于先验分布族 $\Xi = \{N(0, \gamma^2), \gamma > 0\}$ 的渐近最优的 EB 估计 (简记为 a.o. EB 估计).

例 2.20 (单参数离散指数族情形) 设 $P_\theta(X=x):=f_\theta(x)=C(\theta)\theta^x h(x)>0, x=0,1,2,\cdots;\ \theta\in\Theta\subset(0,\infty)$, 损失函数 $L(\theta,a)=(\theta-a)^2$, 参数 θ 的先验分布为 $\xi\in\Xi$, 则 X 的边缘分布为

$$P_\xi^X(X=x):=f_\xi(x)=\int_\Theta f_\theta(x)d\xi(\theta)=h(x)\int_\Theta C(\theta)\theta^x d\xi(\theta). \tag{2.55}$$

因为对所有非负整数 x 有 $f_\xi(x)>0$, 故 θ 在先验分布 ξ 之下的 Bayes 估计为

$$\delta_\xi(x)=\mathrm{E}(\theta|x)=\int_\Theta C(\theta)\theta^{x+1}h(x)d\xi(\theta)/f_\xi(x)=g(x)f_\xi(x+1)/f_\xi(x), \tag{2.56}$$

其中 $g(x)=h(x)/h(x+1)$.

假定 $(X,\theta),(X_1,\theta_1),\cdots,(X_n,\theta_n)$ i.i.d., $X_1,\cdots,X_n$ 和 X 分别为历史样本和当前样本. 对任一非负整数 x, 令

$$\delta_n(x;X^n)=g(x)\Big[\frac{Y_n(x+1)(Y_n(x+1)-1)}{(1+Y_n(x))(2+Y_n(x))}\Big]^{1/2}, \tag{2.57}$$

其中

$$Y_n(x)=X_1,\cdots,X_n \text{ 中等于 } x \text{ 的个数}. \tag{2.58}$$

设先验分布族为

$$\Xi=\Big\{\xi:\mathrm{E}(\theta^2):=\int_\Theta\theta^2 d\xi(\theta)<\infty\Big\}. \tag{2.59}$$

下面我们将证明, 对上述先验分布族, EB 估计 $\delta_n(X;X^n)$ 是渐近最优的. 为此, 首先引进以下引理.

引理 2.18 设 $(Y_1,Y_2,Y_3)\sim$ 多项分布 $M(n;p_1,p_2,p_3)$, 即

$$P(Y_i=y_i,i=1,2,3)=\frac{n!}{y_1!y_2!y_3!}p_1^{y_1}p_2^{y_2}p_3^{y_3},$$

此处 $p_i>0, y_i(i=1,2,3)$ 为非负整数, $\sum_{i=1}^3 p_i=1, \sum_{i=1}^3 y_i=n$, 则有

$$\mathrm{E}\Big[\frac{Y_2(Y_2-1)}{(1+Y_1)(2+Y_1)}\Big]\leqslant\Big(\frac{p_2}{p_1}\Big)^2.$$

证明 分别用 A 和 B 表示非负整数向量 $y=(y_1,y_2,y_3)$ 的如下的集合: $A=\{y:\ y_1+y_2+y_3=n\}$, $B=\{y:\ y_2\geqslant 2, y_1+y_2+y_3=\ n\}$, 则

$$\begin{aligned}
&\mathrm{E}\Big[\frac{Y_2(Y_2-1)}{(1+Y_1)(2+Y_1)}\Big]\\
&=\sum_{y\in A}\frac{n!}{y_1!y_2!y_3!}\frac{y_2(y_2-1)}{(y_1+1)(y_1+2)}p_1^{y_1}p_2^{y_2}p_3^{y_3}\\
&=(p_2/p_1)^2\sum_{y\in B}\frac{n!}{(y_1+2)!(y_2-2)!y_3!}p_1^{y_1+2}p_2^{y_2-2}p_3^{y_3}\leqslant(p_2/p_1)^2.
\end{aligned}$$

证毕. ■

引理 2.19 设 ζ 服从二项分布 $B(n,p)$, 则对任一 $\varepsilon>0$,

$$P(\ |\zeta/n-p|\geqslant\varepsilon)\leqslant 2\exp\Big(-\frac{n\varepsilon^2}{2\lambda+\varepsilon}\Big),$$

此处, $\lambda=\min\{p,1-p\}$.

证明参看 Bennett (1962) 一文中的 Beinstein 不等式.

现在我们来证明下述定理.

定理 2.20 经验 Bayes 估计 $\delta_n(X;X^n)$ 关于 (2.59) 中的先验分布族 Ξ是渐近最优的.

证明 首先证明, 对任一非负整数 x, 有

$$\mathrm{E}[\delta_n(x;X^n)-\delta_\xi(x)]^2\to 0. \tag{2.60}$$

为此, 设 $0<\epsilon<[f_\xi(x)\wedge f_\xi(x+1)]/2$, 且令

$$A_n(x)=\{X^n:|Y_n(x)/n-f_\xi(x)|<\varepsilon\},$$

则在 $A_n(x)\cap A_n(x+1)$ 上, $\delta_n(x;X^n)\leqslant g(x)[f_\xi(x+1)+\varepsilon]/[f_\xi(x)-\epsilon]$, 故存在常数 $C(x)>0$, 使

$$(\delta_n(x;X^n)-\delta_\xi(x))^2I(A_n(x)\cap A_n(x+1))\leqslant C(x).$$

由大数定律, 对 $\forall\, x$, 当 $n\to\infty$ 时, $Y_n(x)/n\to f_\xi(x)$ a.s., 故 $(\delta_n(x;X^n)-\delta_\xi(x))^2\to 0$ a.s., 故由控制收敛定理,

$$\mathrm{E}[\delta_n(x;X^n)-\delta_\xi(x)]^2I\big(A_n(x)\cap A_n(x+1)\big)\to 0. \tag{2.61}$$

另一方面, 由引理 2.19, 不难证明, 存在正的常数 $C(x)$, 使

$$\begin{aligned}&\mathrm{E}[\delta_n(x;X^n)-\delta_\xi(x)]^2\big[I\big(\{A_n(x)\}^C\big)+I\big(\{A_n(x+1)\}^C\big)\big]\\ &\leqslant[2g^2(x)n^2+2\delta_\xi^2(x)]\cdot 4\exp\{-nC(x)\}\to 0.\end{aligned} \tag{2.62}$$

由 (2.61) 和 (2.62), 即得 (2.60).

由引理 2.18,

$$\begin{aligned}\mathrm{E}[\delta_n(x;X^n)-\delta_\xi(x)]^2&\leqslant 2g^2(x)\mathrm{E}\Big[\frac{Y_n(x+1)(Y_n(x+1)-1)}{(1+Y_n(x))(2+Y_n(x))}\Big]+2\delta_\xi^2(x)\\ &\leqslant 2[g(x)f_\xi(x+1)/f_\xi(x)]^2+2\delta_\xi^2(x)=4\delta_\xi^2(x),\end{aligned}$$

由条件 Jensen 不等式,

$$\begin{aligned}\mathrm{E}\big[\delta_\xi(X)\big]^2&=\mathrm{E}\big[\mathrm{E}(\theta|X)\big]^2\\ &\leqslant\mathrm{E}\big[\mathrm{E}(\theta^2|X)\big]=\mathrm{E}(\theta^2)<\infty,\end{aligned}$$

故由 (2.60) 及控制收敛定理, 即得

$$\mathrm{E}[\delta_n(X;X^n)-\delta_\xi(X)]^2\to 0.$$

考虑到 $\mathrm{E}(\theta|X,X^n)=\mathrm{E}(\theta|X)=\delta_\xi(X)$, 由条件期望的平滑性, 易证

$$\mathrm{E}[\delta_n(X;X^n)-\theta]^2-\mathrm{E}[\delta_\xi(X)-\theta]^2=\mathrm{E}[\delta_n(X;X^n)-\delta_\xi(X)]^2\to 0.$$

这就证明了定理 2.20. ■

注 2.2 如果将 (2.57) 中定义的 δ_n 改换为通常的形式:

$$\delta_n(X,X^n)=g(x)\Big[\frac{1+Y_n(x+1)}{1+Y_n(x)}\Big],$$

则引理 2.18 的结论要作如下修改:

$$\mathrm{E}\Big[\frac{1+Y_2}{1+Y_1}\Big]^2 \leqslant 16\Big(\frac{p_2}{p_1}\Big)^2 + 4.$$

由定理证明过程可以看出, 为了证明 δ_n 的渐近最优的性质, 也许还要补充假定 $\mathrm{E}\big[g(X)\big]^2 < \infty$, 不知这一条件在一般情形能否去掉, 即使可以去掉, 看来在数学上也不是一件容易的事. 参看成平, 陈希孺, 陈桂景, 吴传义 (1985) 关于 Poisson 参数的 EB 估计的有关部分.

在例 2.20 中取 $f_\theta(x) = \mathrm{e}^{-\theta}\theta^x/x!, x = 0, 1, 2, \cdots, \theta > 0$, 就得到 Poisson 分布族. 此时, $f_\xi(x) = \int_0^\infty [\mathrm{e}^{-\theta}\theta^x/x!]d\xi(\theta)$, 在平方损失之下 θ 的 Bayes 估计为

$$\delta_\xi(x) = (x+1)f_\xi(x+1)/f_\xi(x).$$

无论是按 (2.57) 还是按注 2.2 定义 $\delta_n(X, X^n)$, 其中, $g(x) = (x+1)$, 则当条件 $\int_0^\infty \theta^2 d\xi(\theta) < \infty$ 成立时, $\delta_n(X, X^n)$ 为 a.o. EB 估计.

例 2.20 排除了二项分布参数的情形. 关于这种情形, 可见下面的例子.

例 2.21 设 $X|\theta \sim B(N, \theta), \theta \in \Theta \subset [0, 1]$, 损失函数 $L(\theta, a) = (\theta - a)^2$, 先验分布为 ξ, 则 θ 的 Bayes 估计为

$$\delta_\xi(x) = \frac{x+1}{N+1}\frac{f_\xi(x+1, N+1)}{f_\xi(x, N)},$$

其中 $f_\xi(x, m) = \int_\Theta \binom{m}{x}\theta^x(1-\theta)^{m-x}d\xi(\theta), x = 0, 1, \cdots, m$, 是当 $X|\theta \sim B(m, \theta), \theta \sim \xi$ 时 X 的边缘分布. 为了作出一个 a.o. EB 估计, 需要通过历史样本 $X_1, \cdots, X_n$ 估计 $f_\xi(x, N)$ 和 $f_\xi(x+1, N+1), x = 0, 1, \cdots, N$. 现 $X_1, \cdots, X_n$ 的共同分布为 $f_\xi(\cdot, N)$, 估计 $f_\xi(\cdot, N)$ 不成问题, 但若事先对先验分布一无所知, 则无法对 $f_\xi(x+1, N+1)$ 作出合用的估计. 但如对先验分布族作某些限制, 例如限制 $\xi \sim Be(a, b), a > 0, b > 0$, 则可以证明 θ 的 a.o. EB 估计存在. 读者如欲知其详, 可参考陈希孺 (1999) 例 3.10 及其所引文献.

2.3.6 关于 Bayes 统计推断的一些说明

如前所述, 在风险函数一致最优准则下, 统计决策问题往往没有最优解. 以统计决策的观点考察 Bayes 估计, 在任何情况下, 一个估计 δ 的 Bayes 风险, 总可视为其风险函数对某个测度 ξ 的加权平均, 以此加权平均作为衡量估计优良性的准则, 在一般情况下, 可以找到一个 δ_ξ, 使此加权平均最小, 这就是 Bayes 估计. 测度 ξ 只是我们选定的一个适宜的加权因子, 不需要有统计意义. 它可以是一概率测度, 也可以是一无限测度.

在某些情形, 有关的统计问题多次反复出现, 各次 θ 的真值遵从某个概率分布 ξ. 以此 ξ 为加权因子, 一个估计 δ 的 Bayes 风险, 即为它在该问题的多次使用中造成的平均损失. 基于此种统计意义, 自然应当选取这个 ξ 为加权因子, 这种先验分布易于为频率学者所接受. 在此情况下, 人们自然要研究如何利用历史资料以提取 ξ 的有用信息, EB 方法就是这样的一种尝试.

在前面的讨论中, 我们已经多少涉及 Bayes 学派的一些观点. 频率学派和 Bayes 学派是当今数理统计学的两大学派. 笼统地说, 凡坚持概率的频率解释, 以及对数理统计学中的概念、方法、结果和性能的评价等, 都是必须在大量重复的意义下去理解的, 属于频率学派; 反之, 则属于 Bayes 学派. 关于这两个学派争论的要点, 在陈希孺, 倪国熙 (1988) 第五章中有精确恰当的介绍, 此处就不再重复了.

下面我们对 Bayes 统计推断作一些简单介绍. 笼统地说, 统计推断是一种探求未知的单纯研究性的工作, 其目的在于获取有关未知参数 θ 的知识, 而不涉及损失或收益等方面的考虑. 那么, Bayes 方法将如何解释? 即如何从统计推断的观点看待 Bayes 方法?

1° 先验分布 ξ 是 Bayes 统计推断的一个要素, 它总结了在得到样本 x 之前, 研究者对 θ 的有关知识或看法.

2° 获取样本 x 的目的, 是对上述知识和看法加以修正和更新, 其结果是把对 θ 的认识由先验分布转化为后验分布 $P(\cdot|x)$. 后验分布综合了 θ 的先验信息和样本 x 包含的 θ 的信息, 后者体现在统计模型即样本分布的表述之中.

按 Bayes 学派的观点, 在获得后验分布 $P(\cdot|x)$ 之后, 统计推断的任务在原则上就完成了. 此时如果为了特定的目的, 需要对 θ 作出某种特定形式的推断, 可以由研究者根据 $P(\cdot|x)$, 以自己认为合适的方式去做. 例如, 如果需要给出 θ 的一个点估计, 研究者可以取 $P(\cdot|x)$ 的期望、中位数, 也可以取后验密度函数的最大值点即众数, 这些都不是 Bayes 方法中固有的, 而只是研究者个人的选择. 由此, 我们可以得出以下结论.

Bayes 统计推断的原则 对参数 θ 的任何推断 (估计、检验等) 只能基于 θ 的后验分布.

此原则可以理解为, 一旦由样本 x 算出了 θ 的后验分布, 就可以设想我们除了此后验分布外, 其余的东西 (样本值、样本分布、先验分布) 全忘记了. 此时, 对 θ 作推断的唯一凭借就是此后验分布. 例如, 无偏性这个规则就不能用, 因为一个估计是否无偏与样本分布族有关, 不是单纯取决于后验分布. 同样的原因, 矩估计也不符合 Bayes 推断原则.

在一定的先验或广义先验分布之下, 极大似然估计 (MLE) 符合 Bayes 推断原则. 事实上, 如选取先验分布为 $d\theta, \theta \in \Theta$, 则易见存在只与样本 x 有关的常数 $c_x > 0$, 使

$$p(\theta|x) = c_x f(x,\theta),\ \theta \in \Theta.$$

此处, $p(\theta|x)$ 为后验密度, $f(x,\theta)$ 为似然函数. θ 的极大似然估计 $\hat{\theta}$, 可由解极值方程 $f(x,\hat{\theta}) = \sup_\theta f(x,\theta)$ 得到, 这与解极值方程 $p(\hat{\theta}|x) = \sup_\theta p(\theta|x)$ 是一回事. 这就是所要的结论. 如果采用一般的先验或广义先验分布, 则 MLE 不一定符合 Bayes 推

断原则. 基于后验密度函数, 我们能做些什么呢? 此时如果存在 $\hat{\theta}$, 使

$$p(\hat{\theta}|x) = \sup_{\theta\in\Theta} p(\theta|x),$$

则可用 $\hat{\theta}$ 去估计 θ, 称为**广义极大似然估计**, 简记为 GMLE. 事实上, 这个估计就是后验密度的众数.

Bayes 推断的思想和方法具有以下一些特点, 使它对应用者有吸引力.

(1) "先验分布 ξ+ 样本 $x \Rightarrow$ 后验分布 $P(\cdot|x)$"这个模式与人们认识事物的过程一致, 即不断以新发现的资料来修正和更新原有的知识或看法.

(2) Bayes 推断在方法上总是落实到 $P(\cdot|x)$ 的计算上, 这可能很复杂但无原则困难, 这是一个固定而不难实现的程式. 在频率学派的方法中, 为了作出推断, 需要知道有关统计量的抽样分布, 这在理论上常常是很难的.

(3) 用 $P(\cdot|x)$ 来表述对未知参数的认识, 显得比单纯用一个统计量表述更自然些. 而对某些常见的问题, Bayes 方法提供的解释, 显得比频率学派提供的解释更合理. 举几个例子.

设 θ 为某事件发生的概率, 研究目的为获取 θ 的知识. 在试验前, 研究者对 θ 没有任何理论或经验的了解. 因此他采用均匀分布 $R(0,1)$ 作为 θ 的先验分布. 以 X 记 n 次重复试验中该事件发生的总次数, X 有二项分布 $B(n,\theta), 0<\theta<1$. 据例 2.13, 当$X=x$ 时 θ 的后验分布为 Beta 分布 $Be(x+1, n-x+1)$. 从先验密度到后验密度的变化, 体现了研究者对 "θ 取值" 这个问题看法的变化. 例如, 现在认为 θ 取值于 x/n 附近可能性更大, 而早先则对 (0, 1) 中的 θ 一视同仁. 在这个例子中, 频率学派的方法是用充分统计量 x/n 去估计 θ, 在某种意义下可以算得上是一个好的估计了. 而按 Bayes 方法, 则只是把 x/n 看成是 θ 最大可能的取值, 是各种可能性之一.

现在回到例 2.14, 考虑 θ 的估计, 不引进任何损失, 先验分布 ξ 和样本分布均不加改变. 仍取后验均值为 θ 的 Bayes 估计, 符合 Bayes 推断原则. 在例 2.14 中对这一估计有一个诠释, 形象地体现了后验分布综合了先验信息和样本信息的本质.

例 2.22 在例 1.12 中我们顺便提到了似然原则. 现举一例说明在有关方面两种统计推断思想的差别.

假定为对某事件发生的概率 θ 作出推断, 我们设计了两个试验, 情况如下:

a. 作 n 次重复试验, 该事件发生了 x 次, 似然函数为 $\binom{n}{x}\theta^x(1-\theta)^{n-x}$;

b. 重复试验直到该事件出现 k 次为止, $k \geqslant 1$, 此时试验的总次数为 y, 按负二项分布, 似然函数为 $\binom{y-1}{k-1}\theta^k(1-\theta)^{y-k}$.

假定碰巧在试验结果中 $x=k, y=n$. 此时两个不同的试验, 不同的模型, 导致只差一个常数因子的似然函数. 按似然原则, 对 θ 应得出同样的推断. Bayes 推断满足这一要求, 因为这时只要两者取同样的先验分布, 不论这先验取何种分布, 后验分布都是一样的, 而 Bayes 推断只能基于后验分布. 但按频率学派惯用的方法, 则推断结果会因模型不同而不同. 例如, 若用 UMVUE 估计 θ, 对模型 a, 所得估计为

$\delta_a(x)=x/n$; 对模型 b, 所得估计为 $\delta_b(y)=I(y=1)$ 或 $\delta_b(y)=(k-1)/(y-1)$, 视 $k=1$ 或 $k\geqslant 2$ 而定. 在 $x=k,y=n$ 时得出不同的估计值.

例 2.23 某种元件的寿命服从指数分布, 其密度为 $\theta\mathrm{e}^{-\theta x}I(x>0),0<\theta<\infty$. 对 n 个元件的寿命作在时刻 T 删失的观测试验, 得样本 $x_1,\cdots,x_n$, ξ 为先验分布, 要据此对 θ 作推断.

若无删失一说, 即样本为完全样本, x_i 的分布为 $F_\theta(x)=(1-\mathrm{e}^{-\theta x})I(x>0)$. 但因事先有删失的约定, x_i 的分布成为 $F_\theta^*(x)=F_\theta(x)$, 当 $x<T$, $F_\theta^*(x)=1$, 若 $x\geqslant T$.

现设这 n 个元件都在时刻 T 前就失效了, 数据事实上虽未被删失, 但因试验条件中有删失一说, 使研究者不能不按后者定样本分布. 按频率学派的方法, 基于后者的推断比基于前者的推断要复杂得多. 一个抱实用观点的人可能会争辩, 既然事实上未被删失, 就该按照前一模型去做.

但若按 Bayes 方法, 则只要样本事实上未被删失, 则不论从 $x_i\sim F_\theta$ 还是 F_θ^* 出发, θ 的后验分布都是一样的, 即为 $c(x^n)\theta^n\exp\{-\theta\sum_{i=1}^n x_i\}d\xi(\theta)I(\theta>0)$, 其中 $c(x^n)=1/\int_0^\infty\theta^n\exp\{-\theta\sum_{i=1}^n x_i\}d\xi(\theta)$.

上面的两个例子引自陈希孺 (1999). 关于截尾和删失的定义和区分, 在 Kendall 和 Stuart (1979) 一书中有精确恰当的叙述, 读者可参看该书的第 2 卷, 32.15—32.22 小节. 简而言之, 删失是关于样本的一个性质, 截尾则是关于分布的一个性质.

(4) 由于统计推断常与客观性诉求相联系, 在 Bayes 推断中采用无信息先验常常是一个容易而合理的分析方法. 参看下面的例子.

例 2.24 设我们碰到的情况是 $X\sim N(\theta,\sigma^2)$, 而 θ 应该是正的. 为简单计可设 σ^2 是已知的. θ 的一个经典的估计是 x, 但当其取负值时显然不合适. 假定无先验信息可用, 一个合理的方法是采用无信息先验密度 $p(\theta)=I(0<\theta<\infty)$. 得出的后验密度为

$$p(\theta|x)=\exp\{-(\theta-x)^2/2\sigma^2\}I(0<\theta<\infty)\Big/\int_0^\infty\exp\{-(\theta-x)^2/2\sigma^2\}d\theta.$$

作变量替换 $\theta=x+\sigma\eta$, 则得后验均值

$$\begin{aligned}\mathrm{E}(\theta|x)&=\int_0^\infty\theta p(\theta|x)d\theta=\frac{\int_{-x/\sigma}^\infty(x+\sigma\eta)\exp\{-\eta^2/2\}\sigma d\eta}{\int_{-x/\sigma}^\infty\exp\{-\eta^2/2\}\sigma d\eta}\\&=x+(2\pi)^{-1/2}\sigma\exp\{-x^2/2\sigma^2\}/(1-\Phi(x/\sigma)).\end{aligned}$$

此处 Φ 是标准正态分布函数.

该例引自 Berger (1985). 该例及其处理方法容易推广到受限制的参数空间的一般情形. 其中一种情况涉及方差分量的估计. 另一个例子是向量参数 $\theta=(\theta_1,\cdots,\theta_k)$ 的分量之间有某种形式的次序限制的情况. 用类似的方法容易得到符合相应要求的估计.

2.3.7 先验分布的选取, 无信息先验分布

用 Bayes 方法处理统计决策和统计推断问题的前提是给出 θ 的先验分布. 因此, 如何选取先验分布是至关重要的问题. 根据前面的介绍, **使用共轭先验分布**显然是一种选择: 一方面可以达到方便数学处理的目的, 另一方面, 由于在这样的族中一般都包含有多种分布, 可以在一定程度上满足统计建模的需要. 下面介绍一些其他可能的选择.

客观法 问题具有在讨论 EB 方法时提到的那种模式: 历史资料提供了 θ 取值的信息, 可用以建立先验分布. 这种做法仍是基于概率的频率解释, 但易于为实用工作者所接受.

主观法 依研究者主观的看法去选择一个先验分布. 按 Bayes 学派的说法, 这是一种通过 "自我反省" 去确定先验分布的方法. 也就是说, 对于参数 θ 取某些值的可能性有多大, 通过思考, 觉得该如何, 而定下一个值. 这种主观先验应当是反映了个人以往对 θ 的了解.

无信息先验 (noninformative priors) 在 Bayes 统计的框架下, 即使在无先验信息可以利用的情形, 仍然需要一个先验分布. 从 Bayes 统计问世之日起, 不少学者就致力于研究在这种情况下如何使用 Bayes 方法. 为此而引进的先验分布统称为无信息先验分布. 无信息先验分布是很难用确切的方式加以定义的, 因此对同一问题不同作者往往从不同的角度, 提出不同的无信息先验. 简单地说, 不包含参数 θ 的任何信息的先验分布就称为一个无信息先验分布. 这只有在 θ 仅取有限个值的简单情形容易说清楚. 例如, 在参数 θ 的两个值之间选择一个的检验问题中, 给每个可能值赋予 1/2 概率的先验显然就是无信息先验. 这样, 我们姑且把无信息先验理解为所有参数值均为同等可能的先验分布. 下面仅就常用的一些无信息先验分几类加以说明, 遗漏之处在所难免.

同等无知原则 这一原则有时称为 Bayes 假定. 作为无信息先验分布, 是 Bayes (1763) 和 Laplace (1812) 引进和使用过的. 以废品率 θ 为例. 正如我们在例 2.13 中那样, 当我们对 θ 一无所知时, 我们只好先验地认为, θ 以同等机会取 (0, 1) 区间内各种值, 因而可取 (0, 1) 区间上的均匀分布 $R(0,1)$ 作为 θ 的先验分布. 这就是所谓同等无知原则.

这个原则有一个困难, 就例 2.13 而言, 如果我们对 θ 一无所知, 则对 θ^3 也一无所知. 按同等无知的原则, 也可以取 $R(0,1)$ 作为 θ^3 的先验分布. 此时 θ 的先验分布就不再是 $R(0,1)$ 了.

下面是另一个例子.

例 2.25 设 $X|\theta \sim N(\theta,1), \theta \in \Theta = (-\infty,\infty)$. 问题是如何选取一个无信息先验分布. 从 "同等无知" 的直观出发, 一个合理的先验密度 $\omega(\theta)$ 在 $(-\infty,\infty)$ 上似应取常数值 $c>0$, c 的选取并不重要, 因而可取 $\omega(\theta)\equiv 1$, 这是一个广义先验分布.

现在我们采用新的参数 $\zeta = \exp\{\theta\}$, 仍设 θ 的无信息先验密度为常数, 则由 $d\theta = d\log\zeta = d\zeta/\zeta$ 可知, $\zeta \sim I(0 < \zeta < \infty)d\zeta/\zeta$. 在这个例子中, 为了在改换参数前后保持问题的一致性, 就得采用这样两个不同的无信息先验. 此时如果要求 θ 和 ζ 的无信息先验密度都为常数, 则会失去这种一致性.

为了避免上面提到的困难, 设想在某种情况下可以在统计模型中引进一个合适的变换群, 使得原来的 (X, θ) 问题和变换后的 (Y, ζ) 问题具有相同的结构, 此时我们就可以选取变换前后保持不变的无信息先验分布. 这种无信息先验可以有严格的定义, 但它只能用于某些特殊类型的分布族的参数. 通过考虑有关的变换群以导出无信息先验的努力始于 Jeffreys, 且已为许多学者广泛使用. Berger (1985) 的第 3.3.2 节列出了有关的参考文献. 下面通过几个例子来说明这种方法.

例 2.26 (位置参数的无信息先验分布) 设样本空间 $\mathcal{X}$ 和参数空间 Θ 都是实数集 $\mathbf{R}$, $X|\theta \sim I(-\infty < x < \infty)f(x-\theta)dx, -\infty < \theta < \infty$. 这样的分布族称为位置参数族, θ 称为位置参数.

为了导出一个无信息先验, 设想将度量原点由 0 移至 $-c$, $c \in \mathbf{R}$, 原来的测量值 X 变为 $Y = X + c$, 参数也由 θ 变为 $\zeta = \theta + c$, 参数族保持不变, 而 Y 有密度 $f(y-\zeta)$. 此时 (X, θ) 问题和 (Y, ζ) 问题具有完全相同的结构, 一个合理的要求是两者有相同的无信息先验密度 $\omega(\cdot)$. 由 $\zeta = \theta + c$ 和密度变换的公式, $\omega(\zeta) = \omega(\zeta - c)$ 对 $\forall\zeta$ 和 $\forall c$ 成立, 这样的无信息先验密度 $\omega(\theta)$ 只能取常数. 这是一个广义先验分布.

例 2.27 (刻度参数的无信息先验分布) 设样本 X 的分布族为参数 θ 的一维密度族

$$I(-\infty < x < \infty)\theta^{-1}f(x/\theta)dx, \quad 0 < \theta < \infty.$$

这样的分布族称为刻度参数族, θ 称为刻度参数.

为了导出一个无信息先验, 设想将测量单位加以改变, 原来的测量值 X 变为 $Y = cX$, 参数也由 θ 变为 $\zeta = c\theta$, 参数族保持不变, 而 Y 有密度 $\zeta^{-1}f(y/\zeta)$. 此时 (X, θ) 问题和 (Y, ζ) 问题具有完全相同的结构, 因而可以要求两者有相同的无信息先验密度 $\omega(\cdot)$. 由 $\zeta = c\theta$, $\omega(\zeta) = c^{-1}\omega(\zeta/c)$ 对 $\forall\zeta$ 和 $\forall c$ 成立, 而 $\omega(\theta) \propto 1/\theta$. 这也是一个广义先验分布.

在上面两个例子中, 实际上我们分别在样本空间 $(\mathbf{R}^1, \mathcal{B}_1)$ 上引进了一个平移变换群和一个刻度变换群 G, 由此导出了参数空间 Θ 上的变换群 $\bar{G}$, 使得 (X, θ) 问题在变换群下是不变的, 这样, 无信息先验就可以在变换前后保持不变的原则下推演出来. 我们有时也会遇到在更一般的机制下, 如何寻求变换群下的不变测度的有关问题. 在某些统计问题中, 存在一个自然的选取, 那就是根据有关情况分别选取左或右不变 Haar 测度. 例如下面例 2.29 中的无信息先验, 也可通过左不变 Haar 测度的简单计算来确定. 关于这方面的一般性讨论, 可参看 Berger (1985) 的第 6 章. 关于在局部紧拓扑群上的 Haar 测度的一般理论, 可参看 Halmos (1950) (王建华译, 1965) 的第 11 章. 下面我们对此做一简单介绍.

局部紧拓扑群上的 Haar 测度

一个拓扑空间称为紧的, 如果空间的任何开覆盖都存在有限的子覆盖. 一个拓扑空间称为局部紧的, 如果空间中的每一点都有一个邻域, 其闭包是紧的. 一个拓扑空间称为 Hausdorff 空间, 如果其中任意两个不同的点, 皆存在互不相交的邻域.

一个群 G 称为是一个拓扑群, 如果它是一个 Hausdorff 空间, 且映照 $(g_1, g_2) \mapsto g_1^{-1}g_2$ 在乘积拓扑空间 $G \times G$ 上是连续的. 此处我们假定涉及的拓扑群总是局部紧的. 此时由所有紧集生成的 σ-域 $\mathcal{B}_G$ 中的每个集合都称为 Borel 集. $\mathcal{B}_G$ 上定义的一个测度称为 Borel 测度, 如果每个紧集都有有限的测度. 一个测度称为是正则的, 如果每个 Borel 集的测度、外测度和内测度都相等.

设局部紧拓扑群 G 上的一个 Borel 测度 μ 具有下列性质: 对于每个非空 Borel 开集 U, $\mu(U) > 0$, 且对于每个 Borel 集 E, $\mu(xE) = \mu(E)$; 则称 μ 是一个左 Haar 测度.

上述定义中的第二个条件通常称为左不变性, 即左转移下的不变性. 同样可以定义一个右 Haar 测度 ν, 它是 G 上的一个 Borel 测度, 赋予每个非空开集以正的测度, 且具有以下的右不变性: 对于每一个 Borel 集 E, $\nu(Ex) = \nu(E)$. 不难验证, 如果 μ 是一个左 Haar 测度, 对于每个 Borel 集 E, 令 $\nu(E) = \mu(E^{-1})$, 此处 $E^{-1} := \{g^{-1} : g \in E\}$, 则 ν 即为一个右 Haar 测度.

关于 Haar 测度的存在性, 我们不加证明地引述如下定理:

在每个局部紧拓扑群 G 上皆存在正则的 (左或右) Haar 测度. 两个不同的 Haar 测度仅相差一个常数倍数.

例 2.28 设 $\bar{G}$ 是由所有形如

$$\begin{pmatrix} 1 & 0 \\ \mu & \sigma \end{pmatrix}$$

的矩阵关于矩阵乘法组成的仿射群, 其中 $-\infty < \mu < \infty, 0 < \sigma < \infty$. 如果将 $\bar{G}$ 看作欧氏半平面而赋予通常的拓扑, 则 $\bar{G}$ 是一个局部紧拓扑群. 对于 $\bar{G}$ 中的每个 Borel 集 E, 令

$$\mu(E) = \iint_E \sigma^{-2} d\mu d\sigma, \qquad \nu(E) = \iint_E \sigma^{-1} d\mu d\sigma,$$

则 μ 和 ν 分别是 $\bar{G}$ 上的左 Haar 测度和右 Haar 测度.

在例 2.10 中若记 $\theta_1 = \mu, \theta_2 = \sigma$, 则参数空间中的变换群即为上述仿射群.

在陈希孺, 倪国熙 (1988) 一书中只把在某种变换群下保持不变的这种先验称为无信息先验.

基于 Fisher 信息矩阵的方法 这是 Jeffreys (1961) 提出的选取无信息先验的另一种方法, 即在参数 θ 为 k 维向量的情形, 选取 $\omega(\theta) = [\det I(\theta)]^{1/2}$ 作为一个无信息先验密度, 其中 $I(\theta)$ 即为样本的 Fisher 信息矩阵. 设 $X \sim f(x|\theta)d\nu(x), \theta \in \Theta = \mathbf{R}^k$, 则

$$I(\theta)=\mathrm{E}_\theta\Big[\frac{\partial l(X,\theta)}{\partial\theta}\frac{\partial l(X,\theta)}{\partial\theta^\tau}\Big],$$

此处 $l(X,\theta)=\log f(X|\theta)$ 为 log 似然函数.

当 $f(x|\theta)$ 对 θ 的各个分量 θ_i 有二阶连续偏导数时, 方阵 $I(\theta)$ 的 (i,j) 元也可表达为

$$I_{i,j}(\theta)=-\mathrm{E}_\theta\Big[\frac{\partial^2 l(X,\theta)}{\partial\theta_i\partial\theta_j}\Big].$$

例 2.29 (位置 − 刻度参数族) 设 $X|(\mu,\sigma)\sim\sigma^{-1}f\big((x-\mu)/\sigma\big)dx$, 其中 $\mu\in\mathbf{R}^1$ 和 $\sigma>0$ 是未知参数, 这样的分布族称为位置-刻度族. 为简单计, 以 $N(\mu,\sigma^2)$ 为例, $\theta=(\mu,\sigma)^\tau$, 不计常数项, 有 $l(X,\theta)=-\log\sigma-(X-\mu)^2/(2\sigma^2)$.

$$I(\theta)=-\mathrm{E}_\theta\begin{pmatrix}\partial^2 l(X,\theta)/\partial\mu^2 & \partial^2 l(X,\theta)/\partial\mu\partial\sigma\\ \partial^2 l(X,\theta)/\partial\mu\partial\sigma & \partial^2 l(X,\theta)/\partial\sigma^2\end{pmatrix}=\begin{pmatrix}1/\sigma^2 & 0\\ 0 & 2/\sigma^2\end{pmatrix},$$

因而 $\omega(\theta)\propto\sigma^{-2}$, 即无信息先验分布为 $\sigma^{-2}d\mu d\sigma$, 这一无信息先验也可通过与例 2.26 和例 2.27 中类似的论证得到. 如果仿照例 2.10 中那样引进有关的变换群, 则它是仿射群 $\bar G$ (参看例 2.28)上的左不变 Haar 测度. 但 Jeffreys (1961) 还是推荐了 $\sigma^{-1}d\mu d\sigma$ 为无信息先验, 它是仿射群 $\bar G$ 上的右不变 Haar 测度.

在例 2.10 所示的正态总体中, 记 $\theta_1=\mu,\theta_2=\sigma$. 容易验证, 如果采用仿射群 $\bar G$ 上的右不变 Haar 测度作为先验分布, 则在例 2.10 的损失之下, (μ,σ) 的 Bayes 估计与最优同变估计相同. 关于更一般的结果, 可参看 Berger (1985) 的第 6.6.2 节. 这也许就是 Jeffreys (1961) 推荐右不变 Haar 测度作为无信息先验的一个原因.

前面提到, 由于无信息先验的界定的不确定性, 基于不同的考虑, 对于同一问题不同作者往往提出不同的无信息先验. 下面举一个简单的例子.

例 2.30 设 $X|\theta\sim B(n,\theta)$, 即

$$P_\theta(X=x)=\binom{n}{x}\theta^x(1-\theta)^{n-x},\quad x=0,1,\cdots,n.$$

参数空间 $\Theta=[0,1]$. 关于 θ 的无信息先验密度, 文献中至少已有以下 4 种:

$\omega_1(\theta)=1$,

$\omega_2(\theta)=\theta^{-1}(1-\theta)^{-1}$,

$\omega_3(\theta)=\theta^{-1/2}(1-\theta)^{-1/2}$,

$\omega_4(\theta)=\theta^\theta(1-\theta)^{1-\theta}$.

参看 Berger (1985) 的第 3.3.4 节.

2.4 Minimax 估计

前面多次提到, 为了发展基于风险函数的统计决策的优化理论, 我们可以提出一些整体性的优良性准则, 以期在这些准则之下可以找到有关问题的最优解. Minimax 准则就是这样的一个准则.

考虑一个一般的统计决策问题, 以 $R(\theta,\delta)$ 记决策函数 δ 的风险函数. 记 $M(\delta)=\sup_{\theta\in\Theta}R(\theta,\delta)$. 如果存在某个 δ_0, 使得

$$M(\delta_0)\leqslant M(\delta),\quad \text{对所有的 } \delta, \tag{2.63}$$

则称 δ_0 为一个**极小化极大** (minimax) **决策函数**; 在估计参数函数 $g(\theta)$ 的决策问题中, 则称 δ_0 为 $g(\theta)$ 的一个**极小化极大** (minimax) **估计**.

为简便起见, 在本节和 2.5 节, 当提及 $g(\theta)$ 的某种估计时, 我们常常会略去 "$g(\theta)$ 的" 这一限定词.

在 minimax 估计的定义中, $M(\delta)$ 是单一指标, 故一般说来 minimax 估计存在, 但具体求解并不容易, 只有在少数简单情形能够做到. 目前讨论得比较多的求解方法一般有两种: 一种是基于 Bayes 方法, 另一种是研究在何种情形最优同变估计为 minimax 估计. 本书只介绍第一种方法.

我们还要指出, 由推论 1.32, 在定理 1.30 的条件下, 对凸损失驱动的 minimax 问题, 只需在基于充分统计量的非随机化估计中寻求 minimax 估计.

定理 2.21 *设 δ_ξ 是在某狭义先验分布 ξ 之下 $g(\theta)$ 的 Bayes 估计, 其风险函数 $R(\theta,\delta_\xi)$ 在参数空间 Θ 上恒等于一个常数 c, 则 δ 为 $g(\theta)$ 的 minimax 估计. 另外, 若 δ_ξ 是先验分布 ξ 之下唯一的 Bayes 估计, 则它也是唯一的 minimax 估计.*

证明 设 δ 是任取的另一估计, 则

$$M(\delta)\geqslant\int_\Theta R(\theta,\delta)d\xi\geqslant\int_\Theta R(\theta,\delta_\xi)d\xi=M(\delta_\xi).$$

这就证明了第一个结论. 又若 δ_ξ 是先验分布 ξ 之下唯一的 Bayes 估计, 则上面的第二个 "$\geqslant$" 为 "$>$" 代替, 因而 δ_ξ 即为唯一的 minimax 估计. ■

定理 2.22 *设 δ_{ξ_n} 是狭义先验分布 ξ_n 之下的 Bayes 估计, $n=1,2,\cdots$, 而估计量 δ 满足条件*

$$\infty\ >\ M(\delta)\leqslant\limsup_{n\to\infty}R_{\xi_n}(\delta_{\xi_n}):=c\leqslant\infty, \tag{2.64}$$

则 δ 为 minimax 估计, 此处 $R_{\xi_n}(\delta_{\xi_n})$ 是 δ_{ξ_n} 的 Bayes 风险.

证明 假定存在 δ^*, 使 $M(\delta^*)<M(\delta)$, 则据 (2.64), 存在自然数 N, 使 $R_{\xi_N}(\delta_{\xi_N})>M(\delta^*)$, 由此可得

$$R_{\xi_N}(\delta^*)\leqslant M(\delta^*)<R_{\xi_N}(\delta_{\xi_N}),$$

与 δ_{ξ_N} 为先验分布 ξ_N 之下的 Bayes 估计矛盾. 定理证毕. ■

这些定理的作用更多的在于验证从其他途径获得的估计是 minimax 估计. 对定理 2.22 而言, 主要的困难在于一串先验分布如何选取, 以及对涉及的 $M(\delta)$ 与 $R_{\xi_n}(\delta_n)$ 如何估值. 此处我们顺便指出, 定理 2.22 不能提供有关 minimax 估计的唯一性, 即使与每个 ξ_n 对应的 Bayes 估计都是唯一的.

容许性与 minimax 性质之间的联系

由于估计的 minimax 性质和容许性之间存在一些内在的联系, 因而我们常常要

把它们放在一起加以讨论. 根据 1.3.3 节容许性的一般定义, 对于参数函数 $g(\theta)$ 的一个估计 δ, 如果存在一个估计 δ^*, 其风险函数严格地一致优于 δ, 则称 δ 为非容许的, 否则就是容许的. 显而易见, 如果一个问题中的 minimax 估计是唯一的, 则它必为容许的, 因为比一个 minimax 估计更好的估计也是 minimax 估计. 这跟一个 UMVU 估计 (或一个 MRE 估计) 的情形不同, 因为比这个 UMVU 估计 (或 MRE 估计) 更好的估计可能是一个有偏估计 (或非同变估计).

由定理 2.21, 我们有如下推论.

推论 2.23 *如果在某个狭义先验分布之下的 Bayes 估计是唯一的, 且具有常数风险, 则必为容许的 minimax 估计.*

关于容许性与 minimax 性质之间的联系, 还有以下的定理.

定理 2.24 *设 δ 是 $g(\theta)$ 的一个估计, 具有常数风险 c 且是容许的, 则 δ 为 $g(\theta)$ 的 minimax 估计. 又若损失函数是严格凸的, 则 δ 为唯一的 minimax 估计.*

证明 设若 δ 不具有 minimax 性质, 即存在另一估计 δ_0, 使 $M(\delta_0) < M(\delta) = c$, 则由 δ 具有常数风险 c, 对任一 $\theta \in \Theta$, 有 $R(\theta, \delta_0) < R(\theta, \delta)$, 与 δ 的容许性假设相矛盾, 这就证明了第一个结论.

今设损失函数是严格凸的, 且 δ_0 为另一 minimax 估计, 则对所有 θ,

$$R\big(\theta, (\delta + \delta_0)/2\big) < \big(R(\theta,\ \delta) + R(\theta,\ \delta_0)\big)/2 = c = R(\theta,\ \delta),$$

与 δ 为容许估计的假定矛盾, 这表明 δ 为唯一的 minimax 估计. ■

例 2.31 设样本 X 服从二项分布 $B(n, \theta), 0 \leqslant \theta \leqslant 1$, 取平方损失 $(\theta - a)^2$, 要估计 θ. 先验分布取为 $Be(a, b)$, 由例 2.13, Bayes 估计 $\delta_{a,b}(X) = (X + a)/(n + a + b)$, 且当取 $a = b = \sqrt{n}/2$ 时其风险函数为一常数 $1/[4(\sqrt{n} + 1)^2]$. 由定理 2.21 和推论 2.23, $\delta^*(X) := \delta_{\sqrt{n}/2, \sqrt{n}/2}(X) = (X + \sqrt{n}/2)/(n + \sqrt{n})$ 是唯一的 minimax 容许估计.

将 minimax 估计的常数风险与 UMVU 估计 X/n 的风险函数 $\theta(1 - \theta)/n$ 作一比较, 不难看出, 仅当 $|\theta - 1/2| < \sqrt{1 - (1 + n^{-1/2})^{-2}}/2$ 时, minimax 估计才比后者有较小的风险, 且差距最大也不过 $(2\sqrt{n} + 1)/[4(n + \sqrt{n})^2]$; 而当 n 很大时上述区间的长度以及风险的最大差距都接近于 0. 这一比较具有典型意义, 即 minimax 决策是一种保守型的策略, 旨在避免在最坏的情况出现时可能遇到的较大风险, 代价是在通常情况下其表现往往不佳, 参看图 2.1. 图 2.1 中虚线表示例 2.31 中 θ 的 minimax 估计的风险函数, 曲线表示估计 X/n 的风险函数.

今设参数空间取为 $(0, 1)$ 开区间, 而损失函数取为 $L(\theta, a) = c(\theta)(\theta - a)^2$, 其中 $c(\theta) = 1/[\theta(1 - \theta)]$. 在此损失之下, X 具有常数风险且为对应于先验分布 $R(0, 1)$ 的唯一的 Bayes 估计. 由定理 2.21 和推论 2.23, X 是唯一的 minimax 容许估计. 另一方面, $\delta^*(X)$ 的风险等于 $1/[4(1 + \sqrt{n})^2\theta(1 - \theta)]$ 是无界的. 这说明对于损失函数中 $c(\theta)$ 的不同选取, 会导致不同的 minimax 估计.

例 2.32 设 $X^n := (X_1, \cdots, X_n)$ 为从正态总体 $N(\theta, \sigma^2)$ 中抽取的 i.i.d. 样本, $-\infty < \theta < \infty$, σ^2 已知, 要估计 θ, 损失为 $(\theta - a)^2$. 取先验分布 $\xi_k = N(0, k^2), k =$

图 2.1 样本量 $n = 1000$, 纵轴刻度为实际的 4500 倍

$1, 2, \cdots$, 由例 2.14, Bayes 估计 $\delta_{\xi_k}(X^n) = n\sigma^{-2}\bar{X}/(n\sigma^{-2} + k^{-2})$, 其风险函数为

$$R(\theta, \delta_{\xi_k}) = (n\sigma^{-2} + k^{-4}\theta^2)/(n\sigma^{-2} + k^{-2})^2.$$

因为当 $\theta \sim N(0, k^2)$ 时 $\mathrm{E}(\theta^2) = k^2$, 故 δ_{ξ_k} 的 Bayes 风险

$$R_{\xi_k}(\delta_{\xi_k}) = 1/(n\sigma^{-2} + k^{-2}) \to \sigma^2/n, \quad \text{当 } k \to \infty.$$

而 $\bar{X}$ 的风险函数为常数 σ^2/n. 由定理 2.22, $\bar{X}$ 是 minimax 估计.

上述结果也可选用一列先验分布 $\{\xi_n = R(-k, k), k = 1, 2, \cdots\}$ 来证, 但过程要复杂得多.

如果在本例中 σ^2 也未知, 即参数空间为 $\{(\theta, \sigma^2) : -\infty < \theta < \infty, \sigma > 0\}$, 则求解 θ 的 minimax 估计的问题失去意义. 因为对任一估计 $\delta(X^n)$, 由 $\bar{X}$ 为 σ 已知时 θ 的 minimax 估计, 知对任一 $\sigma > 0$,

$$\sup_{\theta} \mathrm{E}_{\theta,\sigma}[\delta(X^n) - \theta]^2 \geqslant \mathrm{E}_{\theta,\sigma}(\bar{X} - \theta)^2 = \sigma^2/n, \tag{2.65}$$

因而

$$M(\delta) = \sup\{\mathrm{E}_{\theta,\sigma}[\delta(X^n) - \theta]^2 : -\infty < \theta < \infty, \sigma > 0\}\} = \infty.$$

因此每个估计都是 minimax 估计. 但是, 若改用损失函数 $(\theta - a)^2/\sigma^2$, 则 minimax 估计仍为 $\bar{X}$. 因为由 (2.65), 对任一估计 $\delta(X^n)$,

$$\sup\{\mathrm{E}_{\theta,\sigma}[\delta(X^n) - \theta]^2/\sigma^2 : -\infty < \theta < \infty\} \geqslant 1/n$$

对任一 σ 成立, 故在损失 $(\theta - a)^2/\sigma^2$ 之下,

$$M(\delta) = \sup\{\mathrm{E}_{\theta,\sigma}[\delta(X^n) - \theta]^2/\sigma^2 : -\infty < \theta < \infty, \sigma > 0\} \geqslant 1/n = M(\bar{X}),$$

这就证明了所要的结果. 这一结果也可用定理 2.22 来证明.

这一结果也顺便说明了, 对于损失函数 $L(\theta, a) = c(\theta)(\theta - a)^2$ 中 $c(\theta)$ 的不同选取, 会导致不同的 minimax 估计.

前面我们指出, 在凸损失之下, 我们常常可以只限于寻求非随机化的 minimax 估计. 但在损失函数为非凸的情形, 可以举出这样的例子, 其 minimax 估计只能是随机化的. 在 Hodges 和 Lehmann (1950) 的例子中, 样本 $X \sim B(n, \theta), 0 \leqslant \theta \leqslant 1$. 他们证得, 当 $0 < r < 1$ 时, 在非凸损失 $|\theta - a|^r$ 之下, θ 的 minimax 估计存在, 且不为非随机化估计. 下面这个更为简单的例子取之于 Lehmann (1983).

例 2.33 (随机化的 minimax 估计) 设样本 $X \sim B(n,\theta), 0 \leqslant \theta \leqslant 1$. 为估计 θ, 损失函数取为

$$L(\theta,a)=\begin{cases}0, & \text{当 } |\theta-a| \leqslant \alpha,\\ 1, & \text{否则,}\end{cases}$$

其中 $0<\alpha<1/[2(n+2)]$. 易见, 任何非随机化的 $\delta(x)$ 至多能取 [0,1] 区间中的 $n+1$ 个不同的值, 这些点至多把 [0,1] 区间分隔成 $n+2$ 个子区间, 其中必有一个长度不小于 $1/(n+2)$, 把该区间的中点记为 θ_0. 易见, θ_0 与任一 $\delta(x)$ 的距离都大于 α, 因而 $L(\theta_0,\delta(\cdot))\equiv 1$, $\sup_{0\leqslant\theta\leqslant 1} R(\theta,\delta)=1$. 现取 θ 的一个不依赖于样本 x 的估计 $U\sim U(0,1)$, 其风险函数为 $R(\theta,U)=1-P(|U-\theta|\leqslant\alpha)$, 易见 $\sup_{0\leqslant\theta\leqslant 1} R(\theta,U)=1-\alpha<1$, 因而 θ 的 minimax 估计只可能是随机化的. 下面证明随机化的 minimax 估计存在. 事实上, 对于 θ 的任一随机化估计 $\delta(x,\cdot)$, $F_i(t):=\delta(i,[0,t)), 0\leqslant t\leqslant 1$ 均为 [0,1] 区间上的一个分布函数, 因而这一随机化估计即可表示为 $F=(F_0,\cdots,F_n)^\tau$ 的形式. 此时 δ 的风险函数可表示为

$$R(\theta,F)=\sum_{i=0}^{n}\binom{n}{i}\theta^i(1-\theta)^{n-i}\big(1-F_i(I_\theta)\big)\leqslant 1,\ I_\theta=[\theta-\alpha,\theta+\alpha]\cap[0,1]. \tag{2.66}$$

今设 $\inf_F\sup_\theta R(\theta,F)=v$, 则存在一列 F^k, 使 $\sup_\theta R(\theta,F^k)\to v$. 由 Helly 定理, 存在 F^k 的一个子列 (不妨仍记为 F^k) 和 $F^0=(F_0^0,\cdots,F_n^0)^\tau$, F^0 的每个分量均为 [0,1] 区间上的一个分布函数, 且当 $k\to\infty$ 时, F^k 的每个分量依分布收敛于 F^0 的对应分量. 由弱收敛的等价定义及 I_θ 为闭集, $\limsup_{k\to\infty} F_i^k(I_\theta)\leqslant F_i^0(I_\theta), i=0,1,\cdots,n$. 由 (2.66), $\liminf_{k\to\infty} R(\theta,F^k)\geqslant R(\theta,F^0)$, $v=\lim_{k\to\infty}\sup_\theta R(\theta,F^k)\geqslant\sup_\theta R(\theta,F^0)$, 由此可知, F^0 即为 θ 的一个随机化的 minimax 估计.

2.5 估计的容许性

根据 1.3.3 节容许性的一般定义, 对于参数函数 $g(\theta)$ 的一个估计 δ, 如果存在一个估计 δ^*, 其风险函数严格地一致优于 δ, 则**称 δ 为非容许的, 否则就是容许的**. 在有些情况下, 我们还要考虑一个估计在某个估计类中的容许性, 其定义自然也是不言自明. 本节将对估计的容许性作一些初步的介绍.

涉及容许性的问题有两类: 一是判定一个估计是不是容许的; 一是寻求一个估计问题中的全部容许估计. 从决策理论的角度来看, 我们自然希望采用的估计是容许的, 这就引起了对常用估计的容许性研究的兴趣. 可以想象过去认为优良的估计应当大多是容许的. 但 Stein (1956)、James 和 Stein (1961) 证明了, 在平方损失下估计三维以上正态的均值, 样本均值不是容许的; 随后 Stein (1961) 又证明了, 在平方损失下估计一维正态的方差, 样本方差的线性函数也不是容许的. 这些出人意料的结果, 引起了人们更大的兴趣. 应当指出, 判定一个估计容许与否, 并非容易的事, 至今结果尚不多. 还有, 认为一个容许估计就必然好的观念也是错误的. 例如, 假定

损失函数 $L(\theta,a)$ 满足如下条件：对每个 $\theta\in\Theta$, 存在唯一的 $a(\theta)\in\mathcal{A}$, 使得 $L(\theta,a)$ 在该点取得最小值, 且分布族有共同支撑. 此时若对某个 $\theta_0\in\Theta$, 取 $\delta(x)\equiv a(\theta_0)$, 它显然是 θ 的容许估计, 但由于这个估计只顾及 θ_0 一点而不顾及其余, 因而不会为人们所采用.

先看几个例子.

例 2.34 设 $X_1,\cdots,X_n$ 为从正态总体 $N(\theta,1)$ 中抽取的 i.i.d. 样本, 参数空间为 $|\theta|\leqslant 1$, 要在平方损失 $L(\theta,a)=(\theta-a)^2$ 下估计 θ. 前面已证得, $\bar{X}$ 是 θ 的 UMVU 估计, 但这个估计显然不是容许的, 读者不难写出它的一个改进估计.

例 2.35 设 $X_1,\cdots,X_r$ 相互独立, $X_i\sim N(\mu_i,\sigma^2)$, $\mu_i\in(-\infty,\infty)$ 未知, $i=1,\cdots,r$, $\sigma>0$ 已知. 记 $X=(X_1,\cdots,X_r)^\tau$, $\mu=(\mu_1,\cdots,\mu_r)^\tau$. 损失函数取为 $L(\mu,a)=||\mu-a||^2=\sum_{i=1}^r(\mu_i-a_i)^2$, 其中 $a=(a_1,\cdots,a_r)^\tau$. 下面我们要证明, 当 $r\geqslant 3$ 时, 通常的 UMVU 估计 X 不是容许的. 为此, 考虑 James-Stein 估计

$$\delta_c(X)=\Big(1-\frac{c\sigma^2(r-2)}{||X||^2}\Big)X, \tag{2.67}$$

其中, $0<c<2$ 为一常数. 下面将证明, δ_c 的风险函数 $R(\mu,\delta_c)$一致地小于 X 的风险函数 $r\sigma^2$.

首先, 设 $\delta(x)=x-g(x)$, 其中 $g(x)=(g_1(x),\cdots,g_r(x))^\tau$, 则由分部积分,

$$\begin{aligned}&\int(x_1-\mu_1)g_1(x)\exp\Big\{-\frac{(x_1-\mu_1)^2}{2\sigma^2}\Big\}dx_1\\ =\ &\sigma^2\int\frac{\partial g_1(x)}{\partial x_1}\exp\Big\{-\frac{(x_1-\mu_1)^2}{2\sigma^2}\Big\}dx_1,\end{aligned}$$

故

$$R(\mu,\delta)=\mathrm{E}_\mu||\mu-\delta(X)||^2=r\sigma^2+\mathrm{E}_\mu||g(X)||^2-2\sigma^2\sum_{i=1}^r\mathrm{E}_\mu\Big\{\frac{\partial g_i(X)}{\partial x_i}\Big\}.$$

取 $g(x)=c\sigma^2(r-2)x/||x||^2$, 则 $\delta_c(x)=x-g(x)$, 且 $||g(x)||^2=c^2\sigma^4(r-2)^2/||x||^2$,

$$2\sigma^2\sum_{i=1}^r\frac{\partial g_i(x)}{\partial x_i}=2c\sigma^4(r-2)\sum_{i=1}^r\frac{||x||^2-2x_i^2}{||x||^4}=2c\sigma^4\frac{(r-2)^2}{||x||^2},$$

故

$$0\leqslant R(\mu,\delta_c)=r\sigma^2-c(2-c)\sigma^4(r-2)^2\mathrm{E}_\mu(1/||X||^2)<r\sigma^2=R(\mu,X). \tag{2.68}$$

这就证明了所要的结论.

注 2.3 首先我们指出, 由推论 1.17, 若 $J\sim\mathrm{Poisson}(||\mu||^2/2\sigma^2)$, 且 $\zeta|J=j\sim\chi^2_{2j+r}$, 则 $\zeta\sim\chi^2_{r,\nu}$, 此处 $\nu=||\mu||/\sigma$. 由于当 $X\sim N(\mu,\sigma^2I_r)$ 时, $\sigma^2/||X||^2\overset{d.}{=}1/\zeta$, 利用上述表示, 容易将 $R(\mu,X)-R(\mu,\delta_c)$ 表达为一个级数的形式. 顺便还有 $\mathrm{E}_\mu(\sigma^2/||X||^2)\leqslant\mathrm{E}_0(\sigma^2/||X||^2)=1/(r-2)$.

例 2.36 设样本 X 的均值为 θ, 方差为 $\sigma^2>0$. 在平方损失 $(\theta-a)^2$ 之下, 考虑形如 $\alpha X+\beta$ 的线性估计在线性估计类中的容许性, 则在 (i) $\alpha>1$; (ii) $\alpha<0$; (iii)

$\alpha=1,\beta\neq 0$ 三种情形, $\alpha X+\beta$ 都不是容许的.

事实上, 若以 $R(\theta,\delta)$ 记估计 δ 的风险函数, 则

$$R(\theta,\alpha X+\beta)=\mathrm{E}_\theta(\alpha X+\beta-\theta)^2=\alpha^2\sigma^2+[(\alpha-1)\theta+\beta]^2.$$

现以情形 (ii) 为例, 当 $\alpha<0$ 时, $(\alpha-1)^2>1$, 故

$$\begin{aligned}R(\theta,\alpha X+\beta)&>[(\alpha-1)\theta+\beta]^2=(\alpha-1)^2[\theta+\beta/(\alpha-1)]^2\\&>[\theta+\beta/(\alpha-1)]^2=R(\theta,-\beta/(\alpha-1)),\end{aligned}$$

即此时 $\delta(x)\equiv-\beta/(\alpha-1)$ 为 $\alpha X+\beta$ 的一个改进估计. 其他情形更容易证明.

有关容许性的正面结果, 大都较为繁杂, 且只有在平方损失的场合才有较系统的结果, 参看成平, 陈希孺, 陈桂景, 吴传义 (1985) 的第八章. 本书只涉及一些简单的内容, 可用于建立一些常用估计的容许性. 对于这类问题的证明, 此处只介绍两种方法：一种是应用 Bayes 方法作为工具; 另一种则是在指数族情形应用 C-R 不等式, 正如在稍后的 Karlin 定理的证明中所做的那样.

下面考虑的容许性都是有关估计在所有估计的类中的容许性. 首先讨论 Bayes 方法在这方面的应用.

定理 2.25 *设参数空间 $(\Theta,\mathcal{B}_\Theta)$ 是欧氏的, 对给定的损失函数, δ_ξ 为先验分布 ξ 之下的 Bayes 估计. 假定以下三个条件满足, 则 δ_ξ 是容许估计.*

1° *任何估计的风险函数, 作为参数 θ 的函数, 在 Θ 上处处连续;*

2° *存在一个估计 δ_0, 其 Bayes 风险 $R_\xi(\delta_0)<\infty$;*

3° *Θ 中的每个点都是先验分布 ξ 的支撑点.*

证明 因 δ_ξ 是 Bayes 估计, 由条件 2°, $R_\xi(\delta_\xi)<\infty$. 设 δ_ξ 不是容许估计, 则存在估计 δ, 使 $R(\theta,\delta)\leqslant R(\theta,\delta_\xi)$, 且不等号至少对某个 $\theta_0\in\Theta$ 成立. 由风险函数的连续性假定, 存在 $\varepsilon>0$, 使当 $\theta\in\Theta_0=(\theta_0-\varepsilon,\theta_0+\varepsilon)\cap\Theta$ 时, $R(\theta,\delta)<R(\theta,\delta_\xi)$, 且由条件 3°, $\xi(\Theta_0)>0$. 故

$$\infty>R_\xi(\delta_\xi)=\int_\Theta R(\theta,\delta_\xi)d\xi(\theta)>\int_\Theta R(\theta,\delta)d\xi(\theta)=R_\xi(\delta),$$

这与 δ_ξ 为先验分布 ξ 之下的 Bayes 估计矛盾. ■

注 2.4 证明完全适用于 ξ 为广义先验分布的情形.

例 2.37 设 $X_1,\cdots,X_n$ 为总体 X 的 i.i.d. 样本, $X\sim N(\theta,\sigma^2),-\infty<\theta<\infty$, σ^2 已知, 要估计 θ, 损失函数为 $(\theta-a)^2$. 由指数族的性质, 任一估计的风险函数都是 θ 的连续函数. 由上述定理, 作为一个广义 Bayes 估计, $\bar X$ 为容许估计. 今证对任何常数 $c,d,0<c<1$, $c\bar X+d$ 是 θ 的容许估计. 为此只需选取先验分布 $\xi=N(\nu,\gamma^2)$, 使

$$\frac{n\sigma^{-2}}{n\sigma^{-2}+\gamma^{-2}}=c,\quad \frac{\gamma^{-2}}{n\sigma^{-2}+\gamma^{-2}}\nu=d,$$

则由例 2.14, $c\bar X+d$ 即为先验分布 ξ 之下的 Bayes 估计, 由上述定理, $c\bar X+d$ 是 θ 的容许估计.

不难证明, 结论对 $c=0$ 及 σ^2 未知的情形都成立.

类似地可以证明, 若样本 $X\sim B(n,\theta), 0<\theta<1$, 损失函数为 $(\theta-a)^2$, 则当 $c>0, d>0, c+d<1$ 时 $\delta(X)=cX/n+d$ 为 θ 的容许估计. 同样可证, 广义 Bayes 估计 X/n 也是容许的. 这些都可作为初学者的练习.

下面介绍用于建立估计容许性的另一种方法.

设样本 X 取自单参数指数族,

$$X\sim dP_\theta(x)=f(x,\theta)d\mu(x)=C(\theta)\exp\{\theta T(x)\}d\mu(x),\quad \theta, T \text{ 取实数值},$$

其中 $-\infty\leqslant\underline{\theta}<\theta<\bar{\theta}\leqslant\infty$, $(\underline{\theta},\ \bar{\theta})$ 为自然参数空间. 在平方损失 $L(\theta,a):=\big(g(\theta)-a\big)^2$ 下, Karlin (1958) 研究了 $g(\theta):=\mathrm{E}_\theta(T)$ 的线性估计在何种条件下为容许估计的问题. 正如例 2.36 所示, 形如 $\alpha T+\beta$ 的线性估计当 $\alpha<0$ 和 $\alpha>1$ 时都不是容许的, 而当 $\alpha=0$ 时有关估计为一常数. 因此只需考虑 $0<\alpha\leqslant 1$ 的情形, 这相当于考虑形如

$$\delta_{k,\lambda}(T)=(T+k\lambda)/(1+\lambda) \tag{2.69}$$

的估计, 其中 k,λ 均为常数, 而 $0\leqslant\lambda<\infty$. Karlin 建立了如下的定理.

定理 2.26 (Karlin 定理) *在上述假定下, 若存在 $\theta_0\in(\underline{\theta},\bar{\theta})$, 使*

$$\int_{\underline{\theta}}^{\theta_0}C^{-\lambda}(\theta)\exp(-k\lambda\theta)d\theta=\int_{\theta_0}^{\bar{\theta}}C^{-\lambda}(\theta)\exp(-k\lambda\theta)d\theta=\infty, \tag{2.70}$$

则当 $0\leqslant\lambda<\infty$ 时 (2.69) 中的估计 $\delta_{k,\lambda}$ 是容许的.

对于 $(\underline{\theta},\bar{\theta})=(-\infty,\infty)$ 的情形, 只要 $k=\lambda=0$, 条件 (2.70) 即可满足. 于是在平方损失之下, T 就是 $\mathrm{E}_\theta(T)$ 的容许估计. 这一特例是在 Girshick 和 Savage (1951) 中得到的.

证明 由充分性原则, 只需考虑 $g(\theta)$ 的基于充分统计量 T 的估计. 以 $I(\theta)$ 记分布族的 Fisher 信息函数. 由指数族的性质,

$$g(\theta)=\mathrm{E}_\theta(T)=-d\log C(\theta)/d\theta=-C'(\theta)/C(\theta),$$

$$\sigma^2(\theta):=\mathrm{Var}_\theta(T)=g'(\theta)=\mathrm{E}_\theta[\partial\log f(X,\theta)/\partial\theta]^2=I(\theta).$$

假定存在 $g(\theta)$ 的一个估计 $\delta:=\delta(T)$, 使

$$R(\theta,\delta)\leqslant R(\theta,\delta_{k,\lambda})=\big[\sigma^2(\theta)+\lambda^2\big(g(\theta)-k\big)^2\big]/(1+\lambda)^2,\ \text{对}\ \forall\,\theta, \tag{2.71}$$

则为证定理结论, 只需证明在每个 P_θ 之下, $\delta(T)=\delta_{k,\lambda}(T)$ a.s. 成立.

分别以 $b(\theta)$ 和 $b_0(\theta)$ 记估计 δ 和 $\delta_{k,\lambda}$ 的偏差. 由 C-R 不等式, 对 $\forall\,\theta$,

$$R(\theta,\delta)\geqslant[g'(\theta)+b'(\theta)]^2/I(\theta)+b^2(\theta)=[\sigma^2(\theta)+b'(\theta)]^2/\sigma^2(\theta)+b^2(\theta). \tag{2.72}$$

由以上两式, 对 $\forall\,\theta$,

$$\big[\sigma^2(\theta)+\lambda^2\big(g(\theta)-k\big)^2\big]/(1+\lambda)^2\geqslant b^2(\theta)+[\sigma^2(\theta)+b'(\theta)]^2/\sigma^2(\theta).$$

记 $h(\theta):=b(\theta)-b_0(\theta)=b(\theta)+\lambda\big(g(\theta)-k\big)/(1+\lambda)$, 则上式可改写为

$$\frac{\sigma^2(\theta)+\lambda^2\big(g(\theta)-k\big)^2}{(1+\lambda)^2}$$
$$\geqslant \Big[h(\theta)-\frac{\lambda\big(g(\theta)-k\big)}{1+\lambda}\Big]^2+\Big[h'(\theta)+\frac{\sigma^2(\theta)}{(1+\lambda)}\Big]^2/\sigma^2(\theta),$$

上式可进一步化简为

$$h^2(\theta)+2[h'(\theta)-\lambda(g(\theta)-k)h(\theta)]/(1+\lambda)\leqslant -[h'(\theta)]^2/\sigma^2(\theta)\leqslant 0.$$

再令 $W(\theta)=C^{\lambda}(\theta)\exp(k\lambda\theta)h(\theta)$, 则上式可改写为

$$C^{-\lambda}(\theta)\exp(-k\lambda\theta)W^2(\theta)+2(1+\lambda)^{-1}W'(\theta)\leqslant 0,\quad \theta\in(\underline{\theta},\bar{\theta}). \tag{2.73}$$

由此式及 $\lambda\geqslant 0$ 可推得 $W(\theta)\geqslant 0$ 对所有 θ 成立. 事实上, 若此结论不成立, 即存在 $\theta_0\in(\underline{\theta},\bar{\theta})$, 使 $W(\theta_0)<0$, 则由 $W'\leqslant 0$, $W(\theta)<0$ 对所有 $\theta\geqslant\theta_0$ 成立, 故在区间 $[\theta_0,\bar{\theta})$ 上, (2.73) 可以写成

$$2(1/W(\theta))'=-2W'(\theta)/W^2(\theta)\geqslant(1+\lambda)C^{-\lambda}(\theta)\exp(-k\lambda\theta),$$

对上式两边从 θ_0 到 θ 积分, 即得

$$(1+\lambda)\int_{\theta_0}^{\theta}C^{-\lambda}(u)\exp\{-k\lambda u\}du\leqslant\frac{2}{W(\theta)}-\frac{2}{W(\theta_0)}\leqslant-\frac{2}{W(\theta_0)},$$

此与 $\theta\to\bar{\theta}$ 时上式左方趋于无穷的定理假定相矛盾, 因而 $W(\theta)\geqslant 0$ 对所有 θ 成立.

同样可证对所有 θ, $W(\theta)\leqslant 0$. 因而 $W(\theta)\equiv 0$, 故 $h(\theta)\equiv 0$, $b(\theta)\equiv b_0(\theta)$, 即 $\mathrm{E}_\theta\big(\delta(T)-\delta_{k,\lambda}(T)\big)\equiv 0$, 由 T 的完全性, 在每个 P_θ 之下, $\delta(T)=\delta_{k,\lambda}(T)$ a.s., 如前所说, 这说明 $\delta_{k,\lambda}$ 是容许的. ■

该定理可用于一些常见分布族. 设 $X_1,\cdots,X_n$ 为取自某一维总体的 i.i.d. 样本, 总体分布依次为: 正态分布 $N(\theta,\sigma^2),-\infty<\theta<\infty$, σ^2 已知 或 σ^2 未知; 两点分布 $P_\theta(X=1)=1-P_\theta(X=0)=\theta,\theta\in(0,1)$, 或 $\theta\in[0,1]$; Poisson $(\theta),\theta\in(0,\infty)$ 或 $\theta\in[0,\infty)$. 对上述几种情形, 样本均值 $\bar{X}$ 在平方损失下都是 θ 的容许估计, 有关证明留给读者作为练习.

例 2.38 设 $X_1,\cdots,X_n$ 为从正态总体 $N(0,\sigma^2)$ 中抽取的 i.i.d. 样本, 参数空间为 $\{\sigma>0\}$, 要估计 $2\sigma^2:=1/\omega$, 则 $(X_1,\cdots,X_n)$ 有密度函数 $c\omega^{n/2}\exp(-\omega\sum x_i^2)=C(\theta)\exp[\theta T(x)]$. 此处,

$$\theta=-n\omega/2,\quad C(\theta)\propto(-2\theta/n)^{n/2},\quad T(X)=(2/n)\sum X_i^2.$$

自然参数空间为 $-\infty<\theta<0$.

新参数的引进是为了使 $\mathrm{E}_\theta[T(X)]=1/\omega$. 取平方损失 $[a/(2\sigma^2)-1]^2=(a\omega-1)^2$. 此处我们只寻求形如 $(T+k\lambda)/(1+\lambda),\lambda>0$ 的容许估计. 由 Karlin 定理, 为使上述估计为容许估计, 充分条件 (2.70) 归结为要求对某个常数 $c>0$,

$$\int_{-\infty}^{-c}\mathrm{e}^{-k\lambda\theta}(-2\theta/n)^{-n\lambda/2}d\theta=\int_{-c}^{0}\mathrm{e}^{-k\lambda\theta}(-2\theta/n)^{-n\lambda/2}d\theta=\infty,$$

即

$$\text{(i)}\ \int_c^{\infty}\mathrm{e}^{k\lambda\zeta}\zeta^{-n\lambda/2}d\zeta=\infty\ \text{ 且 }\ \text{(ii)}\ \int_0^{c}\mathrm{e}^{k\lambda\zeta}\zeta^{-n\lambda/2}d\zeta=\infty.$$

易见, 条件 (i) 等价于 $k=0$, 且 $n\lambda/2\leqslant 1$; 或 $k\lambda>0$; 条件 (ii) 等价于 $n\lambda/2\geqslant 1$. 由此得出容许性条件为 (1) $k=0$, 且 $\lambda=2/n$ 或 (2) $\lambda\geqslant 2/n$, 且 $k>0$. 若将 $(T+k\lambda)/(1+\lambda)$ 转换成 $\alpha\sum x_i^2+\beta$ 的形式, 则只要 (a) $\beta=0$ 且 $\alpha=2/(n+2)$ 或 (b) $\beta>0$ 且 $0<\alpha\leqslant 2/(n+2)$ 成立, $\alpha\sum x_i^2+\beta$ 即为 $2\sigma^2$ 的容许估计.

因此, 只要 (a) $\beta=0$ 且 $\alpha=1/(n+2)$ 或 (b) $\beta>0$ 且 $0<\alpha\leqslant 1/(n+2)$ 成立, $\alpha\sum x_i^2+\beta$ 即为 σ^2 的容许估计. 特别地, $\sum_{i=1}^n X_i^2/(n+2)$ 为 σ^2 的容许估计.

例 2.39 今设 $X_1,\cdots,X_n$ 为从正态总体 $N(\mu,\sigma^2)$ 中抽取的 i.i.d. 样本, 参数空间为 $\{\theta=(\mu,\sigma^2):-\infty<\mu<\infty,\sigma>0\}$, 要估计 σ^2, 损失函数为 $L(\theta,a)=(a\sigma^{-2}-1)^2$. 记 $\bar X=\sum_{i=1}^n X_i/n$, $Q_1=\sum_{i=1}^n(X_i-\bar X)^2$, $Q_2=\sum_{i=1}^n X_i^2$. 考察 σ^2 的三个估计: $\delta_1=Q_1/(n-1)$, $\delta_2=Q_1/(n+1)$ 和估计 $\delta_3=Q_2/(n+2)$. 其中 δ_1 为 UMVU 估计, 风险函数为常数 $2/(n-1)$; δ_2 为最优同变估计和 minimax 估计, 风险函数为常数 $2/(n+1)$. 估计 δ_3 的风险函数为 $R(\theta,\delta_3)=2/(n+2)+[2/(n+2)]^2(\mu/\sigma)^4$.

容易验证, 在 σ^2 的形如 αQ_1 的估计类中, $Q_1/(n+1)$ 是风险最小的, 因而 UMVU 估计 $Q_1/(n-1)$ 不是容许的.

由前例结论类推, 也许可以期望在正态总体均值未知的情形, $Q_1/(n+1)$ 应为 σ^2 的容许估计, 但 Stein (1964) 指出, 估计 $\min\{Q_1/(n+1),Q_2/(n+2)\}$ 具有比 $Q_1/(n+1)$ (一致)更小的风险. 参看成平, 陈希孺, 陈桂景, 吴传义 (1985) 第八章定理 4.1 的系.

估计 δ_3 有一个严重的缺点, 即当 $|\mu/\sigma|\to\infty$ 时, $R(\theta,\delta_3)\to\infty$. 但这个估计却是容许的 (参看成平, 陈希孺, 陈桂景, 吴传义 (1985)).

关于 $Q_1/(n+1)$ 的 minimax 性质可参看成平, 陈希孺, 陈桂景, 吴传义 (1985). 该书的第七章习题 8 断言, σ^α 的基于上述正态样本的 MRE 估计 $C_0(\alpha)S^\alpha$ 也是 σ^α 的 minimax 估计, 其中 $S^2=\sum_{i=1}^n(X_i-\bar X)^2$, 但该处的 $C_0(\alpha)$ 应更正为

$$C_0(\alpha)=\frac{\Gamma\big((n-1+\alpha)/2\big)}{2^{\alpha/2}\Gamma\big((n-1+2\alpha)/2\big)}.$$

这一习题已超出本书的要求. 关于这一最优同变估计的推导, 参看本书 2.2.2 节例 2.10.

注 2.5 假定 $X=(X_1,\cdots,X_n)^\tau$ 有密度函数 $f(x-\theta\mathbb{1}_n)$, 其中 f 已知, $\mathbb{1}_n=(1,\cdots,1)^\tau\in\mathbf{R}^n$, $\theta\in\Theta=(-\infty,\infty)$. 在例 2.8 中, 我们在 $\mathrm{E}_0(X_1^2)<\infty$ 的条件下, 证明了在平移群 (2.20) 和平方损失 $L(\theta,a)=(\theta-a)^2$ 之下, θ 的 MRE 估计即为 Pitman 估计:

$$\delta(X)=X_1-\mathrm{E}_0(X_1|Y)\equiv\frac{\int tf(x-t\mathbb{1}_n)dt}{\int f(x-t\mathbb{1}_n)dt},$$

此处的期望是在参数 $\theta=0$ 之下取的, 而 $Y=(X_2-X_1,\cdots,X_n-X_1)^\tau$ 即为平移群下的极大不变量.

关于单个位置参数的 Pitman 估计在平方损失下的容许性, 读者可参看成平, 陈希孺, 陈桂景, 吴传义 (1985) 第八章第 3 节中所介绍的 Stein (1959) 的有关工作. 根

据这项工作, 若

$$\int_{\mathbf{R}^n}\left\{\frac{\int t^2 f(x-t\mathbb{1}_n)dt}{\int f(x-t\mathbb{1}_n)dt}-\left[\frac{\int t f(x-t\mathbb{1}_n)dt}{\int f(x-t\mathbb{1}_n)dt}\right]^2\right\}^{3/2} f(x)dx<\infty,$$

则在平方损失下, 上述 Pitman 估计 $\delta(X)$ 是 θ 的容许的 minimax 估计.

2.6 习题和补充

2.1 设 X_1, X_2 独立且 $X_1 \sim N(\mu_1, 1), X_2 \sim N(\mu_2, 1), -\infty < \mu_1, \mu_2 < \infty$. 证明 $\theta = \min\{\mu_1, \mu_2\}$ 的无偏估计不存在.

2.2 设 $\mathcal{X} = \{-1, 0, 1, 2, \cdots\}$, 总体分布为

$$P_\theta(X=-1)=\theta, \quad P_\theta(X=x)=(1-\theta)^2\theta^x, \quad x=0,1,2,\cdots,\ 0<\theta<1.$$

从中抽取一个 i.i.d. 样本 $X_1, \cdots, X_n$. 试证: 参数函数 $g(\theta)$ 存在 UMVU 估计的充要条件为, 存在常数 a, c, 使 $g(\theta) = a + c(1-\theta)^2$.

2.3 设 $X_1, \cdots, X_n$ 为从 Poisson 总体 $P(\lambda), \lambda > 0$ 中抽取的 i.i.d. 样本.

(1) 证明 $g(\lambda)$ 存在无偏估计的充要条件是: $g(\lambda)$ 为一收敛级数 $\sum_{k=1}^{\infty} a_k\lambda^k$, $0 < \lambda < \infty$.

(2) 记 $T = \sum_{i=1}^n X_i$, 则 $\mathrm{e}^{-2\lambda}$ 的仅依赖于 T 的无偏估计为 $(1-2/n)^T$. 当 $n=1$ 时它成为 $(-1)^T$. 试评论这个估计量.

2.4 设 $X_1, \cdots, X_n$ 为 i.i.d. Cauchy 变量, $X_i - \theta$ 具有密度 $1/[\pi(1+x^2)]$. 证明其样本均值与单个的观测值具有同样的特征函数. 在 θ 的估计问题中这个结果意味着什么?

2.5 设 $X_1, \cdots, X_n$ 为 i.i.d. 随机变量, 其共同分布为 $\{1, \cdots, \theta\}$ 上的均匀分布, 即

$$P_\theta(X_1 = k) = 1/\theta, \quad \text{对 } k = 1, 2, \cdots, \theta,$$

其中 θ 为一未知的自然数.

(1) 证明: 最大的次序统计量 $X_{(n)}$ 是完全充分统计量.

(2) 求 θ 和 θ^2 的 UMVU 估计.

2.6 设 $X_1, \cdots, X_n$ 为从正态总体 $N(a, \sigma^2)$ 中抽取的 i.i.d. 样本, 其中 $a \in \mathbf{R}, \sigma > 0$ 均为未知.

(1) 求 σ^r 的 UMVU 估计 $(n-1 > -r)$.

(2) 求 a^2/σ^2 的 UMVU 估计. 问 n 应满足什么条件?

2.7 设 $X_1, \cdots, X_n$ 是有相同方差 σ^2 的独立随机变量, X_k 的均值为 $k\theta$, θ, σ 均为未知.

(1) 在 θ 的所有线性无偏估计中寻求一个方差最小的无偏估计 $\hat{\theta}$.

(2) 假定 $X_1, \cdots, X_n$ 均为正态变量. 问 $\hat{\theta}$ 是否为 θ 的 UMVU 估计?

2.8 设 $X_1, \cdots, X_m$ 为从正态总体 $N(a, \sigma^2)$ 中抽取的 i.i.d. 样本, $Y_1, \cdots, Y_n$ 为从正态总体 $N(b, \sigma^2)$ 中抽取的 i.i.d. 样本, 两样本相互独立, a, b, σ^2 均为未知. 求 σ^r

和 $(b-a)/\sigma$ 的 UMVU 估计.

2.9 设 $X_1,\cdots,X_n$ 为从指数分布总体 $E(a,1),-\infty<a<\infty$ 中抽取的 i.i.d. 样本. 对给定的 u, 求 $P_a(X_1>u)$ 的 UMVU 估计.

2.10 设 $X_1,\cdots,X_n$ i.i.d., $X_1\sim E(a,b), a\in\mathbf{R}, b>0$, $X_{(1)}\leqslant X_{(2)}\leqslant\cdots\leqslant X_{(n)}$ 为它的次序统计量. 给定整数 $r\geqslant 2$, 假定我们只能观测到 $\big(X_{(1)},X_{(2)},\cdots,X_{(r)}\big)$. 这就是寿命数据的定数删失模型. 在这一模型中求 a 和 b 的 UMVU 估计.

*2.11 在上述定数删失模型中, 求 $P_{a,b}(X_1>u)$ 的 UMVU 估计.

(1) 写出 $\big(X_{(1)},X_{(2)},\cdots,X_{(r)}\big)$ 的联合密度, 证明 $(X_{(1)},T)$ 为充分统计量, 此处 $T=\sum_{i=1}^{r}\big(X_{(i)}-X_{(1)}\big)+(n-r)\big(X_{(r)}-X_{(1)}\big)$.

(2) 令 $Y_1=X_{(1)}, Y_i=(n-i+1)\big(X_{(i)}-X_{(i-1)}\big), i=2,\cdots,n$. 则 $Y_1,\cdots,Y_n$ 相互独立, $Y_1\sim E(a,b/n), Y_2,\cdots,Y_n$ 同分布, $Y_2\sim E(0,b), 2T/b=Y_2+\cdots+Y_r\sim\chi^2_{2r-2}$. 由 1.6.2 节例 1.13, $(X_{(1)},T)$ 的分布族是完全的.

(3) 验证

$$\delta:=\frac{n-1}{n}I\Big(0\leqslant\frac{u-Y_1}{Y_2}<1\Big)+I(Y_1>u)$$

是被估函数的一个无偏估计.

(4) 若记 $a_+=aI(a>0)$, 证明:

$$\mathrm{E}(\delta|X_{(1)},T)=\frac{n-1}{n}\Big(1-\frac{u-X_{(1)}}{T}\Big)_+^{r-2}I(X_{(1)}\leqslant u)+I(X_{(1)}>u).$$

2.12 设总体分布为 $X\sim\Gamma(a,b)$, 其中自由度 a 已知, 刻度参数 $b>0$ 未知. 设 $X_1,\cdots,X_n$ 为从中抽取的 i.i.d. 样本. 对给定的 u, 求 $P_b(X_1>u)$ 的 UMVU 估计.

2.13 设 $X_1,\cdots,X_n$ 为取自 Pareto 分布 $\alpha\zeta^\alpha x^{-(\alpha+1)}I(x>\zeta)dx$ 的 i.i.d. 样本, 其中 $\zeta>0,\alpha>0$.

(1) 当 ζ 已知时求 α 的 UMVU 估计.

(2) 当 α 已知时求 ζ 的 UMVU 估计.

(3) 求 ζ 和 α 的 UMVU 估计.

2.14 设 G 是一个群, G' 是定义了乘法运算的任一集合, η 是 G 到 G' 内的一个映照, 使 $(xy)\eta=(x\eta)(y\eta)$, 则象 $G\eta$ 对于 G' 中的乘法构成一个群.

[这一习题引自: 贾柯勃逊 (1960) 习题 17 第 2 题.]

2.15 设 $X_1,\cdots,X_n$ 具有联合密度 $f_\theta(x)=f(x_1-\theta,\cdots,x_n-\theta), \theta\in\mathbf{R}$, 其中 f 为一已知函数. 要估计 θ. 变换群取为平移群 (2.20), 损失函数为 $L(\theta,a)=\rho(a-\theta)$. 记 $Y_1=X_1-X_n,\cdots,Y_{n-1}=X_{n-1}-X_n, Y=(Y_1,\cdots,Y_{n-1})$.

(1) 若存在 θ 的一个风险有限的同变估计 $\delta_0(X)$, 且对每个 y, 都存在一个数 $v(y)=v^*(y)$, 使 $\mathrm{E}_0\big[\rho\big(\delta_0(X)-v(y)\big)|y\big]$ 达到最小值, 则 θ 的 MRE 估计存在, 且为 $\delta^*(X)=\delta_0(X)-v^*(Y)$.

(2) 设函数 ρ 为非单调的凸函数, 则 θ 的 MRE 估计存在; 如果 ρ 还是严格凸

的, 则 MRE 估计是唯一的.

[参看 Lehmann 和 Casella (1998) 第三章定理 1.10 及其推论.]

2.16 设 $X_1,\cdots,X_n$ 为从均匀分布族 $\{U(\theta-1/2,\theta+1/2),-\infty<\theta<\infty\}$ 中抽取的 i.i.d. 样本, 要估计 θ. 损失函数为 $L(\theta,a)=(a-\theta)^2$. 利用 Pitman 公式证明, 在平移群 (2.20) 之下 θ 的 MRE 估计为 $(X_{(1)}+X_{(n)})/2$.

2.17 设 $X_1,\cdots,X_n$ i.i.d., $X_1\sim E(\theta,1),\theta\in\mathbf{R}$. 要估计 θ. 损失函数取为 $L(\theta,a)=\rho(\theta-a)$, 其中 ρ 为一非单调的凸函数. 变换群取为平移群 (2.20).

(1) 证明: θ 的 MRE 估计为 $X_{(1)}-v^*$, 其中 v^* 最小化 $\mathrm{E}_0[\rho(X_{(1)}-v)]$.

(2) 寻求在平方损失和绝对值损失之下 θ 的 MRE 估计.

2.18 设 $X_1,\cdots,X_n$ 为从均匀分布族 $\{U(0,\theta),\theta>0\}$ 中抽取的 i.i.d. 样本, 要估计 θ. 损失函数为 $L(\theta,a)=|a/\theta-1|^p,p\geqslant 1$. 变换群取为 $G=\{g_c,c>0\}$, 其中 $g_cx_i=cx_i,i=1,\cdots,n$, $\bar{g}_c(\theta)=c\theta$, $g_c^*a=ca$. 容易验证在变换群 G 之下, 该统计决策问题不变, 且 $Z=(X_1/X_n,\cdots,X_{n-1}/X_n,X_n/|X_n|)^\tau$ 为一极大不变量.

(1) 证明完全充分统计量 $X_{(n)}$ 与 Z 独立, 因而 θ 的 MRE 估计有 $X_{(n)}/w$ 的形式, 其中 w 为一仅与 n 有关的常数.

(2) 证明: 当在损失函数的表达式中分别选取 $p=2$ 和 $p=1$ 时, θ 的 MRE 估计分别为 $[(n+2)/(n+1)]X_{(n)}$ 和 $2^{1/(n+1)}X_{(n)}$.

2.19 设 $X_1,\cdots,X_n$ 为从均匀分布族 $\{U(-\theta,\theta),\theta>0\}$ 中抽取的 i.i.d. 样本, 要估计 θ. 损失函数为 $L(\theta,a)=(a/\theta-1)^2$. 变换群

$$G=\Big\{g_{c,s},c>0,s=(s_1,\cdots,s_n)\in\{-1,\ 1\}^n\Big\},$$

其中 $g_{c,s}x_i=cs_ix_i,i=1,\cdots,n$. 求 θ 的 MRE 估计.

[在该例中, 与不变判决问题关联的变换群是由 $\mathbf{R}^n$ 中的刻度群添加对各个坐标平面的反射变换生成的群.]

2.20 设 $X_1,\cdots,X_n$ i.i.d., $X_1\sim U(\theta_1-\theta_2,\theta_1+\theta_2),\theta_1\in\mathbf{R},\theta_2>0$, 要估计 (θ_1,θ_2). 变换群取为例 2.10 中的仿射群, 损失函数如 (2.30) 式所示, 其中 $r=1$. 求 (θ_1,θ_2) 的 MRE 估计.

2.21 设 $X=(X_1,\cdots,X_n)^\tau$ 为取自分布族 $\{N(0,\sigma^2),\sigma^2>0\}$ 的 i.i.d. 样本, 要估计 σ^r, 其中 $r>0$ 已知. 损失函数为 $L(\sigma,a)=|a/\sigma^r-1|^p$, $p\geqslant 1$. 求在例 2.9 的刻度变换群之下 σ^r 的最优同变估计. 特别地, 写出 $p=2$ 时 σ 和 σ^2 的 MRE 估计的具体形式.

2.22 在例 2.12 中, 仍将先验分布取为 ξ, 而将损失函数取为 $\lambda(\theta)|g(\theta)-a|$, 其中函数 λ 只取正值. 记 $K(a,x)=\mathrm{E}(\lambda(\theta)|\theta-a|\,|x)$. 试根据 $K(\cdot,x)$ 取值的三种不同情形, 分别写出 $g(\theta)$ 的 Bayes 估计.

2.23 将题 2.22 推广到损失为 $\lambda(\theta)\rho_p\big(g(\theta)-a\big)$ 的一般情形, 其中

$$\rho_p(z)=2z[p-I(z<0)].$$

2.24 设随机变量 X 有密度函数

$$f(x|\theta)=\theta \mathrm{e}^{-\theta x} I(x>0), \quad \theta>c,$$

$c>0$ 为已知常数, 要估计 θ. 假定 θ 的先验分布为

$$\pi(\theta)d\theta=\mathrm{e}^{-(\theta-c)} I(\theta>c)d\theta.$$

(1) 试求 θ 的后验密度.

(2) 试求 θ 在绝对值损失 $L(\theta,a)=|\theta-a|$ 之下的 Bayes 估计.

2.25 设样本 X 取自超几何分布

$$P_\theta(X=x)=\binom{\theta}{x}\binom{N-\theta}{n-x}\Big/\binom{N}{n}, \quad x=\max\{0,n-(N-\theta)\},\cdots,\min\{n,\theta\}.$$

此处 N 和 n 已知而 θ 为介于 0 和 N 之间的未知整数, 问题是要在平方损失下估计 θ/N. 取 θ 的离散先验分布 ξ 为集合 $\{0,1,\cdots,N\}$ 上的均匀分布, 求 θ/N 的 Bayes 估计.

2.26 把上题推广到先验分布 ξ 为

$$\xi(\{\theta\})=\frac{\Gamma(2c)}{\Gamma(c)^2}\int_0^1 \binom{N}{\theta} u^{\theta+c-1}(1-u)^{N-\theta+c-1}du, \quad \theta=0,1,\cdots,N$$

的一般情形, 此处 $c\geqslant 1$ 为一实数.

2.27 设 $X_1,\cdots,X_n$ 为取自均匀分布族 $\{R(0,\theta),\theta>0\}$ 的 i.i.d. 样本, 试应用定理 2.15 中的方法, 求与此分布族共轭的先验分布族. 这个分布族正是 Pareto 分布族.

2.28 设 $X_1,\cdots,X_n$ 为取自均匀分布族 $\{R(0,\theta),\theta>0\}$ 的 i.i.d. 样本, θ 的先验分布为 Pareto 分布, 其密度为 $\pi(\theta)=\alpha\zeta^\alpha/\theta^{1+\alpha}I(\theta>\zeta)$, 此处 $\zeta>0,\alpha>0$ 是已知常数. 试求 θ 的后验密度以及 θ 在绝对值损失 $L(\theta,a)=|a-\theta|$ 下的 Bayes 估计.

[所得结果再次表明, Pareto 分布族是与分布族 $\{R(0,\theta)^{*n},\theta>0,n\geqslant 1\}$ 共轭的先验分布族.]

2.29 设 X 服从负二项分布,

$$P_\theta(X=x)=\binom{r+x-1}{x}\theta^r(1-\theta)^x, \quad x=0,1,\cdots,\ 0<\theta<1,$$

$X_1,\cdots,X_n$ 为其 i.i.d. 样本, 试应用定理 2.15 中的方法, 求与此分布族共轭的先验分布族.

2.30 试应用定理 2.15 中的方法, 求与 Poisson 分布序列共轭的先验分布族.

2.31 设样本 X 的分布为 $P_\theta(X=1)=1-P_\theta(X=0)=\theta,0<\theta<1$, 损失函数为 $(a-\theta)^2$, 先验分布为 $P(\theta=\theta_0)=P(\theta=1-\theta_0)=1/2$. 试决定 θ_0, 使在此先验分布之下 θ 的 Bayes 估计为 minimax 估计.

[该例说明, 不同的先验分布可导致相同的 Bayes 解.]

2.32 设 $X\sim P_\theta,0\leqslant\theta\leqslant 1$, 损失函数为 $L(\theta,a)=(1-\theta)a+(1-a)\theta,\theta,a\in[0,1]$. 试求 θ 的基于 X 的一个 minimax 决策函数.

2.33 将一个不均匀的铜板独立投掷三次, 每次出现正面的概率为 $\theta = 1/4$ 或 $\theta = 3/4$, 以 X 记其正面出现的总次数.

(1) 指出 X 的分布类型和参数空间, 求 θ 的极大似然估计 $\hat{\theta}$.
设 $a_1 = 1/4$ 及 $a_2 = 3/4$ 为可采取的行动的全体. 取绝对值损失 $L(\theta, a) = |a - \theta|, a = a_i, i = 1, 2$.

(2) 设 θ 的先验分布 H_p 为 $P(\theta = 1/4) = p = 1 - P(\theta = 3/4)$, 其中 $p \in (0, 1)$ 为给定的数, 试求在先验分布 H_p 之下的 Bayes 判决函数.

(3) 证明 (1) 中给出的 $\hat{\theta}$ 为绝对值损失下 θ 的 minimax 估计.

2.34 设 $X_1, \cdots, X_n$ 为从正态总体 $N(0, \sigma^2)$ 中抽取的 i.i.d. 样本, 参数空间为 $\{\sigma > 0\}$, 损失函数为 $L(\sigma^2, a) = (\sigma^2 - a)^2/\sigma^4$, 则 $\sum_{i=1}^n X_i^2/(n+2)$ 为 σ^2 的 minimax 估计.

2.35 设样本 X 取自超几何分布

$$P_\theta(X = x) = \binom{\theta}{x}\binom{N-\theta}{n-x}\Big/\binom{N}{n}, \quad x = \max\{0, n - (N - \theta)\}, \cdots, \min\{n, \theta\}.$$

此处 N 和 n 已知而 θ 为介于 0 和 N 之间的未知整数, 问题是要在平方损失下估计 θ/N.

(1) 证明: 若取 $\alpha = \{1 + [(N-n)/(nN-n)]^{1/2}\}^{-1}$, $\beta = (1-\alpha)/2$, 则 $\delta(X) = \alpha X/n + \beta$ 的风险函数为一常数.

(2) 证明此 $\delta(X)$ 是 θ/N 的 minimax 估计和在先验分布

$$\xi(\{\theta\}) = \frac{\Gamma(2c)}{\Gamma(c)^2}\int_0^1 \binom{N}{\theta} u^{\theta+c-1}(1-u)^{N-\theta+c-1}du, \quad \theta = 0, 1, \cdots, N$$

下的 Bayes 估计, 此处 $c = \beta/(\alpha/n - 1/N)$.

[选自 Shao (2005) 第 4 章习题 31.]

*2.36 设 $X_1, \cdots, X_n$ 为 i.i.d. $U(0, \theta)$ 变量, $\theta > 0$. 损失函数取为 $L(\theta, a) = (a - \theta)^2/\theta^2$, 求 θ 的 minimax 估计.

[提示: 给定 n, 选取 Pareto(ζ, α) 作为 θ 的先验分布, 求得 Bayes 解, 再令 $\zeta = \alpha \downarrow 0$.]

2.37 设 X 为取自分布族 $\mathcal{P} = \{P_\theta, \theta \in \Theta\}$ 的样本, 要在平方损失下估计 θ. 假定 $\theta_0 \in \Theta$, $\mathcal{P} \ll P_{\theta_0}$, 即 $P_\theta \ll P_{\theta_0}$, 对 $\forall\, \theta \in \Theta$, 则估计 $\delta(X) \equiv \theta_0$ 是容许的.

2.38 设 X 为取自分布族 $\{P_\theta, \theta \in \Theta\}$ 的样本, 当 $\theta \in \Theta$ 时, $g(\theta) \in [c, d] \subset \mathbf{R}$, 其中 c 和 d 均为已知. 假定行动空间包含 $[c, d]$ 且损失函数为 $L(|g(\theta) - a|)$, 其中 $L(\cdot)$ 为 $[0, \infty)$ 上的增函数 (指严增, 下同). 证明: 对任一决策函数 $\delta(X)$, 如果存在 $\theta \in \Theta$, 满足 $P_\theta\big(\delta(X) \notin [c, d]\big) > 0$, 则 $\delta(X)$ 不是容许的.

2.39 设 X 和 Y 为 Bernoulli $(1, p)$ 的独立观测值, $q = 1 - p$. 令 $\theta = g(p) = p^2 + q^2$.

(1) 寻求 θ 的 UMVU 估计 $\hat{\theta}$.

(2) 证明在平方损失下 $\hat{\theta}$ 不是 θ 的容许估计, 并构造它的一个改进估计.

2.40 设样本 X 服从负二项分布,

$$P_p(X=x)=\binom{x-1}{r-1}\Big(\frac{p}{1-p}\Big)^r\exp\{x\log(1-p)\},\quad x=r,r+1,\cdots,$$

r 已知, $0<p<1$未知, 证明在平方损失之下, 当 $0<\alpha\leqslant r/(1+r)$ 且 $\beta\geqslant r(1-\alpha)$ 时, $\alpha X+\beta$ 为 E_pX 的容许估计.

2.41 设 $X_1,\cdots,X_n$ 为取自 θ 的指数分布 $E(0,\theta)$ 的 i.i.d. 样本, $\theta>0$ 为待估的均值, $\bar{X}$ 为样本均值. 证明:

(1) 对任一 $b\geqslant 0$, $(n\bar{X}+b)/(n+1)$ 为平方损失下 θ 的容许估计;

(2) $n\bar{X}/(n+1)$ 为在损失函数 $L(\theta,a)=(a-\theta)^2/\theta^2$ 之下 θ 的 minimax 估计;

(3) $\bar{X}$ 不是容许估计.

2.42 设样本 X 取自 Poisson 分布 $P(\theta),\theta>0$. 证明在损失函数 $L(\theta,a)=(a-\theta)^2/\theta$ 之下, X 为 θ 的容许的 minimax 估计.

2.43 设样本 X 取自 Poisson 分布 $P(\theta),\theta>0$. 试找出 (α,β) 的所有可能的值, 使 $\alpha X+\beta$ 为平方损失下 θ 的容许估计.

[提示: 例 2.36 中提供了一些非容许的情形. 容易证明当 $\beta<0$ 时, 形如 $\alpha X+\beta$ 的估计也是非容许的.]

2.44 设样本 X 取自分布 $[(1-\theta^2)/2]\exp\{x\theta-|x|\}dx,|\theta|<1$. 证明: 当 $0\leqslant\alpha\leqslant 1/2$ 时, 对任意的实数 β, $\alpha X+\beta$ 为 $\mathrm{E}_\theta X=2\theta/(1-\theta^2)$ 在平方损失下的容许估计. [选自 Shao (2005) 第 4 章习题 37.]

2.45 设样本 $X\sim P_\theta:\ P_\theta(X=x)=\theta^x\mathrm{e}^{-\theta}/[x!(1-\mathrm{e}^{-\theta})],\ x=1,2,\cdots,\theta>0$. 考虑在平方损失下对 $\theta/(1-\mathrm{e}^{-\theta})$ 的估计问题.

(1) 证明 X 是容许估计, 且 $\sup_{\theta>0}R(\theta,X)=\infty$;

(2) 试设定一个损失函数使 X 为容许的 minimax 估计.

第三章

假设检验

假设检验是数理统计的一个基础性的分支. 现在通行的假设检验理论体系是 J. Neyman 和 E. S. Pearson 在 20 世纪二三十年代建立和发展起来的, 他们在 1933 年的论文是关于假设检验理论的基本文献. 按照这种理论, 或是后来 Wald 的统计决策理论, 假设检验问题可以归结为一个数学优化问题. 本章将遵循这一线路讨论有关的理论和方法. 鉴于优化问题在多数情形难于求解, 通常的途径是从直观的考虑出发, 提出一些看来合理的检验方法, 再对其统计性质加以研讨. 我们将在后面的章节介绍这些内容.

3.1 基本概念

3.1.1 统计假设和检验函数

设随机元 X 取值于可测空间 $(\mathcal{X}, \mathcal{B})$, 其分布族为 $\{P_\theta, \theta \in \Theta\}$. 一个**统计假设**, 简称**假设**, 是关于 X 的分布即关于参数 θ 的一个命题:

$$H: \ \theta \in \Theta_0,$$

此处, Θ_0 是 Θ 的一个给定的非空真子集. 与数学中常见的"假设函数 f 连续"、"假设 X 的方差有限" 之类的表述不同, 这一命题或陈述, 不是作为一个已被认定的事实, 其正确与否, 要依据试验或抽样结果去作出判断和决定, 即对该假设作出检验.

设参数真值为 θ_0, 则 $\theta_0 \in \Theta_0$, 或 $\theta_0 \notin \Theta_0$, 两者必居其一. 若是前者, 则称为**假设 H 成立**或**假设 H 为真**. 若是后者, 则称 H **不成立**或 H **不真**.

记 $\Theta_1 = \Theta - \Theta_0$, 则命题

$$K:\ \theta \in \Theta_1$$

称为 H 的一个**对立假设**或备择假设. 与此相对, $H:\theta\in\Theta_0$ 常称为**原假设**、零假设或解消假设. 以后我们只将 H 和 K 分别称为原假设和对立假设. 无论是 Θ_0 还是 Θ_1, 如果只含一个参数, 则相应的假设称为**简单假设**, 否则就称为**复合假设**.

把 H 和 K 排在一起写成

$$H:\theta\in\Theta_0 \leftrightarrow\ K:\theta\in\Theta_1 \tag{3.1}$$

或

$$H:\theta\in\Theta_0\ \ \text{vs.}\ \ K:\theta\in\Theta_1$$

的形式, 就称为一个**(统计)假设检验问题**, 简称**检验问题**.

检验一个假设, 就是采取一定的统计程序, 以作出 "接受(承认)" 或者 "否定 (拒绝)" 原假设 H 的结论或行动. 结论的获得不能光凭经验或主观的考虑, 必须基于试验或观测得到的样本 x. 作出决策的过程取决于许多因素, 如样本的分布、检验水平的选定, 以及经验知识的运用 (Bayes 方法) 等, 但最后的决定则取决于样本 x.

在一个检验问题中, 基于样本 x 所作的是否接受 H 的决定, 就称为一个**检验**. 一个**非随机化检验**等同于把样本空间 $\mathcal{X}$ 分成两个互斥子集 $\mathcal{X}_A$ 和 $\mathcal{X}_R$: 当 $x\in\mathcal{X}_A$ 时接受 H, 而当 $x\in\mathcal{X}_R$ 时否定 H. $\mathcal{X}_A$ 和 $\mathcal{X}_R$ 分别称为该检验的**接受域**和**否定 (拒绝) 域**.

更一般地, 可以引进随机化检验程序. 一个**随机化检验**就是一个定义于 $(\mathcal{X},\mathcal{B})$ 取值于 [0, 1] 区间的可测函数 $\phi(x)$. 一旦有了样本 x, 即以 $\phi(x)$ 的概率否定 H, 以 $1-\phi(x)$ 的概率接受 H. 这样的一个 ϕ 称为一个**检验函数**. 若 ϕ 只取 0, 1 两个值, 此时检验的否定域 $\mathcal{X}_R=\{x:\phi(x)=1\}$, 这就回到了非随机化检验的情形. 随机化检验在理论上占有一席之地, 但在实用中并不常用.

3.1.2 假设检验问题的 Neyman-Pearson 提法

设 ϕ 为检验问题 (3.1) 的一个检验函数, 则函数

$$\beta_\phi(\theta)=\int_{\mathcal{X}}\phi(x)dP_\theta(x)=\mathrm{E}_\theta[\phi(X)],\quad \theta\in\Theta$$

称为 ϕ 的**功效函数**. 在假设检验的 Neyman-Pearson 理论中, 功效函数包含了检验函数的全部性质, 功效函数相同的两个检验尽管可以不一样, 但在性质方面没有分别.

设 $0\leqslant\alpha\leqslant 1$. 若

$$\beta_\phi(\theta)\leqslant\alpha,\ \text{对所有}\ \theta\ \in\ \Theta_0,$$

则称 α 为检验 ϕ 的一个水平. 如是, 若 α 为 ϕ 的水平而 $\alpha\leqslant\alpha'\leqslant 1$, 则 α' 也是水平. $\sup_{\theta\in\Theta_0}\beta_\phi(\theta)$ 即为 ϕ 的所有水平的下确界, 即最小水平, 有时称为 ϕ 的**真实水平**. 但通常难以界定所设定的 α 是否为真实水平, 因此当谈及一个检验的水平时, 允许弹性的理解是有益的.

当一个检验 ϕ 用于检验假设 (3.1) 时, 可能犯以下两类错误之一: 一是 H 正确, 但被否定了, 称为**第一类错误**; 一是 H 不真, 但被接受了, 称为**第二类错误**. 当原假设成立时只能犯第一类错误, 而在对立假设之下, 只能犯第二类错误. 以 $E_i(\theta)$ 记使用检验 ϕ 而参数真值为 θ 时犯第 i 类错误的概率, 则有

$$E_1(\theta)=\begin{cases}\beta_\phi(\theta), & \text{若 } \theta\in\Theta_0,\\ 0, & \text{若 } \theta\notin\Theta_0.\end{cases}$$

$$E_2(\theta)=\begin{cases}0, & \text{若 } \theta\in\Theta_0,\\ 1-\beta_\phi(\theta), & \text{若 } \theta\notin\Theta_0.\end{cases}$$

所以, 我们所期望的检验 ϕ, 其功效函数 $\beta_\phi(\theta)$ 应在 Θ_0 上尽可能小, 在 $\Theta_1=\Theta-\Theta_0$ 上尽可能大. 但两者往往不可兼得. 在 Neyman 和 Pearson 构建的理论体系中, 一个重要的原则就是 **"控制第一类错误概率" 的原则**. 这一原则归结为: 要求检验的水平不超过某个给定的数 α, 在这一前提下, 要求功效函数 β_ϕ 在 Θ_1 上尽可能大. 由于这个缘故, β_ϕ 在某个 $\theta\in\Theta_1$ 处的值常称为检验 ϕ 在 θ 点的**功效**.

3.2 一致最优检验

定义 3.1 设 ϕ 是检验问题 (3.1) 的一个水平为 α 的检验, $0\leqslant\alpha\leqslant 1$. 若对 (3.1) 的任一水平为 α 的检验 ψ, 都有

$$\beta_\psi(\theta)\leqslant\beta_\phi(\theta),\ \text{对任一 } \theta\in\Theta_1,$$

则称 ϕ 为 (3.1) 的一个水平为 α 的**一致最优检验** (Uniformly Most Powerful Test), 简记为 **UMP 检验**. 而当对立假设为简单假设时, 这样的检验称为 MP 检验.

按照这一准则, 在符合 "水平为 α" 这一约束的所有检验中, UMP 检验是最好的. 下面我们将研究在什么样的一些检验问题中存在 UMP 检验, 以及怎样寻求 UMP 检验.

3.2.1 Neyman-Pearson 基本引理

为了寻求有关检验问题的给定水平的 UMP 检验, 首先需要引进 Neyman-Pearson (以下简记为 N-P) 基本引理. 为此, 考虑原假设和对立假设都是简单假设的情形, 即考虑如下的检验问题:

$$H: X\sim P_0\leftrightarrow K: X\sim P_1. \tag{3.2}$$

此处, $X=x$ 为样本. 易见 P_0,P_1 总可写成 $dP_i(x)=f_i(x)d\mu(x)$ 的形式, $i=1,2$, 其中 μ 为 $(\mathcal{X},\mathcal{B})$ 上的一个测度. 例如, 可取 $\mu=P_0+P_1$.

不妨设 $f_0(x),f_1(x)$ 均为有限且不同时为 0. 当 $a>0$ 约定 $a/0=+\infty$.

定理 3.1 (N-P 基本引理) 给定 $\alpha\in[0,1]$. 在上述条件下, 有

(i) (存在性) 对检验问题 (3.2), 存在检验 ϕ, 满足条件

$$\phi(x)=\begin{cases}1, & \text{若 } f_1(x)/f_0(x) > k,\\ 0, & \text{若 } f_1(x)/f_0(x) < k,\end{cases} \tag{3.3}$$

其中 k 为某一常数, 且

$$\mathrm{E}_0[\phi(X)]=\int_{\mathcal{X}}\phi(x)f_0(x)d\mu(x)=\alpha. \tag{3.4}$$

(ii) (最优检验的充分条件) 若检验 ϕ 同时满足 (3.3) 和 (3.4), 则 ϕ 必为检验问题 (3.2) 的水平为 α 的 MP 检验.

(iii) (最优检验的必要条件) 若 ϕ 是检验问题 (3.2) 的水平为 α 的 MP 检验, 则对某个常数 k, (3.3) a.e. μ 成立. 又若 ϕ 的功效小于 1, 即

$$\mathrm{E}_1[\phi(X)]=\int_{\mathcal{X}}\phi(x)f_1(x)d\mu(x)<1, \tag{3.5}$$

则 ϕ 也满足 (3.4).

证明 当 $\alpha=0$ 和 $\alpha=1$ 时定理显然成立, 只要允许 k 可取 $+\infty$, 且在积分中将 $0\cdot\infty$ 理解为 0. 以下设 $0<\alpha<1$.

(i) 定义 $h(c)=P_0[f_1(X)/f_0(X)>c], -\infty<c<\infty$, 则 $h(\cdot)$ 非增、右连续且 $h(\infty):=\lim_{c\to\infty}h(c)=0, h(0-)=1$. 故存在 $k\in[0,\infty)$, 使 $h(k)\leqslant\alpha\leqslant h(k-)$. 令

$$\phi(x)=\begin{cases}1, & \text{若 } f_1(x)/f_0(x)>k,\\ \gamma, & \text{若 } f_1(x)/f_0(x)=k,\\ 0, & \text{若 } f_1(x)/f_0(x)<k,\end{cases} \tag{3.6}$$

其中 $\gamma=[\alpha-h(k)]/[h(k-)-h(k)]$, 则 ϕ 同时满足 (3.3) 和 (3.4).

(ii) 设 ϕ 同时满足 (3.3) 和 (3.4), 且 ψ 为检验问题 (3.2) 的一个水平为 α 的检验. 记

$$S^+=\{x:\phi(x)>\psi(x)\},\quad S^-=\{x:\phi(x)<\psi(x)\}, \tag{3.7}$$

则 $x\in S^+ \Rightarrow \phi(x)>0 \Rightarrow f_1(x)/f_0(x)\geqslant k$. 同理, $x\in S^- \Rightarrow f_1(x)/f_0(x)\leqslant k$. 故对任一 x, $[\phi(x)-\psi(x)][f_1(x)-kf_0(x)]\geqslant 0$, 因此

$$\int_{\mathcal{X}}[\phi(x)-\psi(x)][f_1(x)-kf_0(x)]d\mu(x)\geqslant 0,$$

故

$$\int_{\mathcal{X}}[\phi(x)-\psi(x)]f_1(x)d\mu(x)\geqslant k\int_{\mathcal{X}}[\phi(x)-\psi(x)]f_0(x)d\mu(x)\geqslant 0,$$

即 ϕ 为检验问题 (3.2) 的水平为 α 的 MP 检验.

(iii) 设 ψ 是检验问题 (3.2) 的水平为 α 的 MP 检验. 由 (i)—(ii), 存在检验 ϕ 满足 (3.3) 和 (3.4), 且 ϕ 也是 (3.2) 的水平为 α 的 MP 检验. 定义 $S=S^+\cup S^-$, 其中 S^+ 和 S^- 如 (3.7) 所示. 令 $S_1=S\cap\{x:f_1(x)/f_0(x)\neq k\}$, 则在 S_1 上有

$g(x) := [\phi(x) - \psi(x)][f_1(x) - kf_0(x)] > 0$, 故

$$\begin{aligned} 0 \leqslant \int_{S_1} g(x)d\mu &= \int_{\mathcal{X}} g(x)d\mu \\ &= \mathrm{E}_1[\phi(X) - \psi(X)] - k\mathrm{E}_0[\phi(X) - \psi(X)] \\ &= -k\{\alpha - \mathrm{E}_0[\psi(X)]\} \leqslant 0. \end{aligned}$$

故 $\mu(S_1) = 0$, 即 $\phi(x) \neq \psi(x) \Rightarrow f_1(x)/f_0(x) = k$, a.e. μ, 因而至多除去一个 μ-零集, ψ 同 ϕ 一样满足 (3.3).

设 ϕ 为 (3.2) 的水平为 α 的 MP 检验, 且 (3.5) 成立. 今若 $\mathrm{E}_0[\phi(X)] < \alpha$, 令

$$\psi(x) = \min\{1, \phi(x) + \alpha - \beta_\phi(\theta_0)\},$$

则 $\mathrm{E}_0[\psi(X)] \leqslant \alpha$, 即 ψ 是 (3.2) 的水平为 α 的检验, 且对所有 x 有 $\psi(x) \geqslant \phi(x)$, 等号当且仅当 $\phi(x) = 1$ 时成立. 由 $\mathrm{E}_1[\phi(X)] < 1$, 知 $P_1[\phi(X) = 1] < 1$, 由此可知, $\mathrm{E}_1[\psi(X)] > \mathrm{E}_1[\phi(X)]$, 与 ϕ 为 (3.2) 的水平为 α 的 MP 检验矛盾. 故 $\mathrm{E}_0[\phi(X)] = \alpha$, 即 ϕ 也满足 (3.4). 定理证毕. ■

推论 3.2 *设 ϕ 为 (3.2) 的水平为 α 的 MP 检验, 则 $\mathrm{E}_1[\phi(X)] \geqslant \alpha$. 若 $0 < \alpha < 1$ 而 $P_0 \neq P_1$, 则 $\mathrm{E}_1[\phi(X)] > \alpha$.*

证明 以 ϕ 与恒等于 α 的检验比较即得前一结论. 今设 $0 < \alpha < 1$ 而 $P_0 \neq P_1$, 若 $\mathrm{E}_1[\phi(X)] = \alpha$, 则恒等于 α 的检验也是水平为 α 的 MP 检验, 由定理 3.1 的 (iii), 后一检验应有 (3.3) 的形式, 由 $0 < \alpha < 1$ 知 $f_1(x)/f_0(x) = k$, a.e. μ, 其中 k 只能为 1, 因而 $f_1(x) = f_0(x)$, a.e. μ, 这与 $P_0 \neq P_1$ 矛盾. ■

3.2.2 单调似然比分布族与 UMP 检验

N-P 引理虽然只考虑了简单原假设和简单对立假设的情形, 但有其重要的理论意义. 在实际应用中, 原假设和对立假设通常都是复合的. 最简单的情形是, 参数空间是一个一维区间, θ_0 为其内点, 要检验的是一个单边假设: $H : \theta \leqslant \theta_0 \leftrightarrow K : \theta > \theta_0$. 对于一个特定的 $\theta_1 > \theta_0$, 检验问题 $H \leftrightarrow \theta_1$ 的 MP 检验一般会与 θ_1 有关因而不是 UMP 的. 但若添加适当的条件, 则 UMP 检验就会存在. 本节考虑的单调似然比分布族就是这样的一个例子.

定义 3.2 设样本 $X \sim dP_\theta(x) = f(x, \theta)d\mu(x), \theta \in \Theta$. $\Theta \subset \mathbf{R}$ 为一区间. 若存在一个统计量 T, 使对参数空间中任意的 $\theta < \theta'$, 分布 $P_\theta \neq P_{\theta'}$, 且比值 $f(x, \theta')/f(x, \theta)$ 为 $T(x)$ 的非降函数, 则称这样一个分布族为**单调似然比** (Monotone Likelihood Ratio) **分布族**, 简称 **MLR 族**.

定理 3.3 *假定 X 的分布族是一个关于 $T(x)$ 的单调似然比族.*

(i) *设检验函数具有如下形式:*

$$\phi(x) = \begin{cases} 1, & \text{若 } T(x) > c, \\ \gamma, & \text{若 } T(x) = c, \\ 0, & \text{若 } T(x) < c, \end{cases} \tag{3.8}$$

其中 c,γ 为常数, $0\leqslant\gamma\leqslant 1$, 则其功效函数

$$\beta(\theta)=\mathrm{E}_\theta[\phi(X)]$$

非降, 且在集合 $\{\theta\in\Theta:0<\beta(\theta)<1\}$ 上严格增.

(ii) 给定 $\alpha\in(0,1)$, 则可找到形如 (3.8) 的一个检验 ϕ, 满足

$$\mathrm{E}_{\theta_0}[\phi(X)]=\alpha, \tag{3.9}$$

这一检验即为检验问题 $H:\theta\leqslant\theta_0\leftrightarrow K:\theta>\theta_0$ 的水平为 α 的 UMP 检验.

(iii) 对任一 $\theta<\theta_0$, 在满足 (3.9) 式的检验中, 上述检验使 $\beta(\theta)$ (第一类错误的概率) 取最小值.

证明 先证 (i) 和 (ii). 任给 $\theta_0<\theta_1$, 考虑检验问题 $\theta=\theta_0\leftrightarrow\theta=\theta_1$. 令 $h(v):=P_{\theta_0}[T(X)>v]$, 类似定理 3.1 (i), 易证存在 c 和 γ 使 (3.8) 和 (3.9) 成立. 由 MLR 族的定义, $T(x)\geqslant c\Rightarrow f(x,\theta_1)/f(x,\theta_0):=g[T(x)]\geqslant g(c)$. 类似定理 3.1 的 (ii), 所得检验 ϕ 是检验问题 $\theta_0\leftrightarrow\theta_1$ 的水平为 $\alpha=\beta(\theta_0)$ 的 MP 检验. 由推论 3.2 及 $0<\alpha=\beta(\theta_0)<1$, 知 $\beta(\theta_1)>\beta(\theta_0)$. 因 $\theta_0<\theta_1$ 和 $\alpha\in(0,1)$ 都是任取的, (i) 即获证.

由 (i), ϕ 在原假设 $\theta\leqslant\theta_0$ 之下也有水平 α, 因而它是 $\theta\leqslant\theta_0\leftrightarrow\theta_1$ 的水平为 α 的 MP 检验. 由于 ϕ 只依赖于 θ_0 和 α, 不赖于 θ_1, 因而它也是 $\theta\leqslant\theta_0\leftrightarrow\theta>\theta_0$ 的水平为 α 的 UMP 检验. (ii) 也获证.

(iii) 可由考虑下面的对偶问题和检验 $1-\phi$ 的相应性状而获证. ■

关于对偶问题 $H:\theta\geqslant\theta_0\leftrightarrow K:\theta<\theta_0$, 其水平为 α 的 UMP 检验 ϕ 可由将 (3.8) 式中的所有不等式都反过来得到, 即检验 ϕ 具有如下形式:

$$\phi(x)=\begin{cases}1, & \text{若 } T(x)<c,\\ \gamma, & \text{若 } T(x)=c,\\ 0, & \text{若 } T(x)>c,\end{cases}$$

其中 c 和 $\gamma\in[0,1]$ 由

$$\mathrm{E}_{\theta_0}[\phi(X)]=\alpha$$

确定. 读者不难写出定理 3.3 中其余的对偶结论.

将该定理直接用于一维指数族, 可得到如下定理.

定理 3.4 设样本 X 有分布

$$f(x,\theta)d\mu=C(\theta)\exp\{Q(\theta)T(x)\}d\mu,$$

其中, θ 为实参数, 参数空间 Θ 为一非退化区间, θ_0 为 Θ 的内点. 若 $Q(\cdot)$ 在 Θ 上严格增, 则对任给的 $\alpha\in(0,1)$, 检验问题 $\theta\leqslant\theta_0\leftrightarrow\theta>\theta_0$ 的水平为 α 的 UMP 检验 ϕ 有 (3.8) 的形式, 且满足 (3.9). 这个 UMP 检验还满足定理 3.3 的 (iii).

如果 Q 是严格下降函数, 则 (3.8) 中的不等式反向.

例 3.1 本例考虑在下述分布族中, 检验问题 $H:\theta\leqslant\theta_0\leftrightarrow K:\theta>\theta_0$ 的水平为 α 的 UMP 检验.

1° 设 $X_1,\cdots,X_n$ 为从一维正态总体 $\{N(\theta,\sigma^2),\theta\in\mathbf{R}\}$ 中抽取的 i.i.d. 样本, σ^2 已知. 水平为 α 的 UMP 检验的否定域为 $\{\bar X>\theta_0+u_\alpha\sigma/\sqrt{n}\}$, $\bar X=\sum_{i=1}^n X_i/n$, 而 u_α 为标准正态分布的 $(1-\alpha)$ 分位数, 又称上 α 分位数.

2° 设 $X_1,\cdots,X_n$ 为从指数分布总体 $\{E(0,\theta),\theta>0\}$ 中抽取的 i.i.d. 样本. 在分布参数为 θ 时 $2n\bar X/\theta\sim\chi^2_{2n}$. 故上述检验问题的水平为 α 的 UMP 检验有否定域 $\{\bar X>\theta_0\chi^2_{2n}(\alpha)/(2n)\}$, 此处 $\chi^2_{2n}(\alpha)$ 为分布 χ^2_{2n} 的上 α 分位数.

3° 设样本 $X\sim B(n,\theta),0\leqslant\theta\leqslant1$, 水平为 α 的 UMP 检验满足

$$\phi(X)=1,\gamma\text{ ,或 }0,\text{ 如果 }X>c,=c,\text{或 }<c, \tag{3.10}$$

而 c,γ 由条件

$$\sum_{i=c+1}^{n}\binom{n}{i}\theta_0^i(1-\theta_0)^{n-i}+\gamma\binom{n}{c}\theta_0^c(1-\theta_0)^{n-c}=\alpha$$

决定.

4° 设 $X_1,\cdots,X_n$ 为从 Poisson 分布总体 $\{P(\theta),\theta>0\}$ 中抽取的 i.i.d. 样本, 水平为 α 的 UMP 检验满足 (3.10), 而 c,γ 由条件

$$\exp(-n\theta_0)\Big[\sum_{i=0}^{c-1}(n\theta_0)^i/i!+(1-\gamma)(n\theta_0)^c/c!\Big]=1-\alpha$$

决定.

本例考虑的都是单参数指数族. 下面考虑的 MLR 族是非指数族的一个例子.

例 3.2 (超几何分布) 一批产品总数 N 已知, 从中无放回随机抽出 n 个, 总体次品数 M 未知, X 为抽出的次品数. 希望基于 X 检验假设 $H:M\leqslant M_0\leftrightarrow M>M_0$. 样本中的 X 服从超几何分布:

$$P\{X=x\}=P_M(x)=\binom{M}{x}\binom{N-M}{n-x}\Big/\binom{N}{n},$$

其中 $\max\{0,n+M-N\}\leqslant x\leqslant\min\{M,n\}$. 容易验证这个分布族是关于 $T(x)=x$ 的 MLR 族. 由定理 3.3, 水平为 α 的 UMP 检验满足 (3.10), 其中 $T(X)=X$, 而 c,γ 由条件

$$\sum_{x=c+1}^{\min\{M_0,n\}}P_{M_0}(x)+\gamma P_{M_0}(c)=\alpha,\ 0\leqslant\gamma\leqslant1$$

决定.

例 3.3 设 $X_1,\cdots,X_n$ 为从均匀分布族 $\{U(0,\theta),\theta>0\}$ 中抽取的 i.i.d. 样本. 给定 $\theta_0>0,\alpha\in(0,1)$, 要寻求检验问题 $\theta\leqslant\theta_0\leftrightarrow\theta>\theta_0$ 的水平为 α 的 UMP 检验.

任取 $\theta_1>\theta_0$. 先考虑检验问题 $\theta=\theta_0\leftrightarrow\theta=\theta_1$. 记 $x^n=(x_1,\cdots,x_n)^\tau,x_{(n)}=\max\{x_1,\cdots,x_n\}$. 似然比

$$f(x^n,\theta_1)/f(x^n,\theta_0)=\begin{cases}(\theta_0/\theta_1)^n,\text{ 当 }0<x_{(n)}<\theta_0,\\ \infty,\qquad\quad\text{当 }\theta_0\leqslant x_{(n)}<\theta_1.\end{cases}$$

此处, 似然比限于区域 $\{T=x_{(n)}<\theta_1\}$ 中才有意义. 可以把 MLR 族的定义加以扩展, 使其包括现在的情形, 而不会影响定理 3.3 及其证明. 由于似然比在两个分段区

间上分别为常数, 因此满足条件

$$(1)\ \phi(x^n)=1, 若\ x_{(n)}>\theta_0; \qquad (2)\ \mathrm{E}_{\theta_0}[\phi(X^n)]=\alpha$$

的任何检验必为 $\theta=\theta_0 \leftrightarrow \theta=\theta_1$ 的水平为 α 的 MP 检验.

由此可知, 满足条件

$$(1)\ \phi(x^n)=1, 若\ x_{(n)}>\theta_0;\ (2)\ \mathrm{E}_{\theta_0}[\phi(X^n)]=\alpha;\ (3)\ \mathrm{E}_{\theta}[\phi(X^n)]\leqslant\alpha, 若\ \theta<\theta_0$$

的任何检验 ϕ 必为 $\theta\leqslant\theta_0 \leftrightarrow \theta=\theta_1$ 的水平为 α 的 MP 检验. 只要这样的检验与 θ_1 无关, 则必为 $\theta\leqslant\theta_0 \leftrightarrow \theta>\theta_0$ 的水平为 α 的 UMP 检验.

特别地, 可取检验函数 $\phi(x^n)=\psi(x_{(n)})$, ψ 为 $(0,\infty)$ 上的非降函数, 则满足条件

$$(1)\ \psi(x_{(n)})=1, 若\ x_{(n)}>\theta_0; \qquad (2)\ \mathrm{E}_{\theta_0}[\psi(X_{(n)})]=\alpha$$

的任何检验 $\psi(x_{(n)})$ 必为 $\theta\leqslant\theta_0 \leftrightarrow \theta>\theta_0$ 的水平为 α 的 UMP 检验.

这是因为, 若 $\theta<\theta_0$, 则

$$\begin{aligned}\mathrm{E}_{\theta}[\psi(X_{(n)})]&=\int_0^{\theta}\psi(x)d(x/\theta)^n=\int_0^1\psi(t^{1/n}\theta)dt\\&\leqslant\int_0^1\psi(t^{1/n}\theta_0)dt=\mathrm{E}_{\theta_0}[\psi(X_{(n)})]=\alpha.\end{aligned}$$

该检验问题的水平为 α 的 UMP 检验的两个具体的例子为:

(i) 否定域为 $\{X_{(n)}>\theta_0(1-\alpha)^{1/n}\}$ 的检验 ψ_1;

(ii) $\psi_2(x_{(n)})=1$, 或 α, 视 $x_{(n)}\geqslant\ \theta_0$ 或 $x_{(n)}<\theta_0$ 而定.

应当指出, UMP 检验只在一些常用分布族中才存在, 且大多限于单边检验的问题, 至于双边检验的情形, 我们在习题中给出了存在 UMP 检验的几个例子.

3.2.3 假设检验与两决策问题

从某种意义上说, 假设检验可以视为一个二中择一的决策问题, 即行动空间中仅有两个行动 a_0,a_1, 分别表示接受和否定原假设. 在统计决策的框架下, 损失函数可取为 $L(\theta,a_i)=L_i(\theta), i=0,1, \theta\in\Theta$. 有了样本 x 时采取行动 a_1 的(条件)概率 $\phi(x)$ 即为一个随机化决策函数, 其风险函数为

$$R(\theta,\phi)=\mathrm{E}_{\theta}\{L_0(\theta)[1-\phi(X)]+L_1(\theta)\phi(X)\}.$$

问题归结为在 (基于风险函数的) 某种准则下寻求最优的 ϕ.

按这种提法, 一点也不像一个假设检验问题. 若要引进原假设和对立假设, 只要像前面那样, 把参数空间 Θ 分为 Θ_0 和 $\Theta_1=\Theta-\Theta_0$ 两部分即可. 选取损失函数的原则是, 决定对了损失为 0, 决定错了则依哪一类错误及参数值 θ 而定, 即对 $i=0,1$, 若 $\theta\in\Theta_i$, 则取 $L_i(\theta)=0$, 否则取 $L_i(\theta)>0$. 风险为

$$R(\theta,\phi)=\begin{cases}L_1(\theta)\beta_{\phi}(\theta), & 当\ \theta\in\Theta_0,\\ L_0(\theta)[1-\beta_{\phi}(\theta)], & 当\ \theta\in\Theta_1.\end{cases}$$

此处 $\beta_{\phi}(\theta)=\mathrm{E}_{\theta}[\phi(X)]$. 上式中的两个组成部分可以分别视为第一、二类错误的风险. 如要控制第一类错误的风险, 则可以提出一些不同的优化问题. 例如, 可以要求

在约束条件

$$L_1(\theta)\beta_\phi(\theta) \leqslant \alpha, \quad \theta \in \Theta_0$$

之下寻求 ϕ 使 $\sup_{\theta\in\Theta_1} L_0(\theta)[1-\beta_\phi(\theta)]$ 取最大值.

容易看出, 若 $L_0(\theta)$ 在 Θ_1 上为常数, $L_1(\theta)$ 在 Θ_0 上为常数, 则问题归结为假设检验问题. 容许 $L_0(\theta)$ 在 Θ_1 上, $L_1(\theta)$ 在 Θ_0 上与参数 θ 有关的决策问题虽然更加一般和合理, 但因只是提供了一个一般框架, 未能提供多少具体解法, 而无法取代假设检验在数理统计中的地位.

3.3 Neyman-Pearson 基本引理的推广

在本节中, 我们要将 N-P 基本引理推广到更一般的情形, 用以解决检验函数在多个积分等式约束下的极值问题. 如前所述, 一个检验函数就是一个定义在 $(\mathcal{X},\mathcal{B})$ 上取值于 [0,1] 区间的可测函数.

为达此目的, 我们需要引用附录 A.1 节中建立的检验函数空间的一个弱紧性定理. 设 μ 为样本空间 $(\mathcal{X},\mathcal{B})$ 上的一个 σ-有限测度. 称一列检验函数 ϕ_n 弱收敛于 ϕ, 若对任一 μ-可积函数 p,

$$\int \phi_n(x)p(x)\ d\mu(x) \to \int \phi(x)p(x)\ d\mu(x),$$

记为 $\phi_n \overset{w.}{\to} \phi$. 称检验函数空间的一个子集 $\mathcal{G}$ 是弱紧的, 若 $\mathcal{G}$ 中的任一列函数都有一个弱收敛的子列, 且极限函数仍在 $\mathcal{G}$ 中. 我们有如下定理.

定理 3.5 设 $f_1,\cdots,f_{m+1}$ 是欧氏空间 $(\mathcal{X},\mathcal{B})$上的 μ-可积实值函数. 用 $\mathcal{C}$ 表示满足条件

$$\int \phi f_i d\mu = c_i,\ i=1,\cdots,m \tag{3.11}$$

的检验函数 ϕ 的集合. 假定 $\mathcal{C}$ 非空, 则有

(i) 在 $\mathcal{C}$ 中存在一个检验函数, 取得积分 $\int \phi f_{m+1}d\mu$ 在 $\mathcal{C}$ 中的最大值.

(ii) 函数 $\phi\in\mathcal{C}$ 取得上述最大值的一个充分条件为, 存在常数 $k_1,\cdots,k_m$, 使

$$\phi(x) = \begin{cases} 1, & \text{如果}\ \ f_{m+1}(x) > \sum_{i=1}^m k_i f_i(x), \\ 0, & \text{如果}\ \ f_{m+1}(x) < \sum_{i=1}^m k_i f_i(x). \end{cases} \tag{3.12}$$

(iii) 若 $\mathcal{C}$ 中的检验函数 ϕ 满足 (3.12) 且 $k_1,\cdots,k_m \geqslant 0$, 则 ϕ 在所有满足

$$\int \phi f_i d\mu \leqslant c_i,\ i=1,\cdots,m \tag{3.13}$$

的检验函数中使积分 $\int \phi f_{m+1}d\mu$ 取得最大值.

(iv) 设 M 为由形如 $(\int \phi f_1 d\mu,\cdots,\int \phi f_m d\mu)^\tau$ 的 m-维向量组成的集合, 其中 ϕ 遍历所有的检验函数, 则 M 是 $\mathbf{R}^m$ 中的闭凸集. 若 $(c_1,\cdots,c_m)^\tau$ 为 M 在 $\mathbf{R}^m$ 中的

内点, 则存在常数 $k_1,\cdots,k_m$ 和一个检验函数 ϕ 满足 (3.11) 和 (3.12), 而 $\mathcal{C}$ 中的一个函数取得 $\int\phi f_{m+1}d\mu$ 的最大值的一个必要条件是 (3.12) a.e. μ 成立.

(v) 若存在常数 $k_1,\cdots,k_m$, 使 $\mathcal{C}$ 中的检验函数 ϕ 满足

$$\phi(x)=\begin{cases}1, & \text{如果}\quad f_{m+1}(x)<\sum\limits_{i=1}^{m}k_if_i(x),\\ 0, & \text{如果}\quad f_{m+1}(x)>\sum\limits_{i=1}^{m}k_if_i(x).\end{cases}$$

则 ϕ 使积分 $\int\phi f_{m+1}d\mu$ 取得在 $\mathcal{C}$ 中的最小值.

证明 (i) 设 $\{\phi_n\}$ 为 $\mathcal{C}$ 中的一列检验函数使 $\int\phi_n f_{m+1}d\mu\to\sup_{\phi\in\mathcal{C}}\int\phi f_{m+1}d\mu$. 由检验函数的弱紧性定理, 存在一个子列 $\{\phi_{n'}\}$ 和一个检验函数 ϕ^*, 使

$$\int\phi_{n'}f_id\mu\to\int\phi^*f_id\mu,\ i=1,\cdots,m+1.$$

易见 $\int\phi^*f_{m+1}d\mu=\sup_{\phi\in\mathcal{C}}\int\phi f_{m+1}d\mu$, 且 $\phi^*\in\mathcal{C}$.

(ii) 和 (iii) 的证明类似于定理 3.1 (ii) 的证明.

(iv) M 是闭集可由弱紧性定理得到. M 的凸性显然.

设 N 为由形如 $(\int\phi f_1d\mu,\cdots,\int\phi f_{m+1}d\mu)^\tau$ 的 $(m+1)$-维向量组成的集合, 其中 ϕ 遍历所有的检验函数, 则同样可证 N 为闭凸集. M 和 N 中的点分别用 $(u_1,\cdots,u_m)^\tau$ 和 $(u_1,\cdots,u_{m+1})^\tau$ 表示. 分两种情况证明.

1° N 在 $\mathbf{R}^{m+1}$ 中没有内点. 因 N 为包含原点的凸集, 故 N 中的所有的点都在过原点的一个超平面上, 这个超平面具有 $u_{m+1}=\sum_{i=1}^{n}k_iu_i$ 的形式, 故对所有的检验函数 ϕ,

$$\int\phi f_{m+1}d\mu=\sum_{i=1}^{m}k_i\int\phi f_id\mu,$$

这蕴含了 $f_{m+1}(x)=\sum_{i=1}^{m}k_if_i(x)$, a.e. μ, 此时 (3.12) 自然成立.

2° N 在 $\mathbf{R}^{m+1}$ 中有内点. 则由 $(c_1,\cdots,c_m)^\tau$ 为 M 的内点及 N 为闭凸集, N 中前 m 个坐标为 $c_1,\cdots,c_m$ 的点的第 $(m+1)$ 个坐标充满一个非退化的闭区间 $[c_*,c^*],c_*<c^*$. 点 $(c_1,\cdots,c_m,c^*)^\tau$ 为 N 的边界点, 由凸集的性质, 存在过该点的一个超平面 $\Pi:\sum_{i=1}^{m+1}k_iu_i=\sum_{i=1}^{m}k_ic_i+k_{m+1}c^*$, 使 N 中的所有的点都在 Π 的下方或在超平面 Π 上. 只要 $c_*<c<c^*$, N 的内点 $(c_1,\cdots,c_m,c)^\tau$ 就不在超平面 Π 上, 故 $k_{m+1}\neq0$, 取 $k_{m+1}=-1$, 则 N 中的点满足

$$u_{m+1}-\sum_{i=1}^{m}k_iu_i\leqslant c^*-\sum_{i=1}^{m}k_ic_i,$$

故对所有的检验函数 ϕ,

$$\int\Big(f_{m+1}-\sum_{i=1}^{m}k_if_i\Big)\phi d\mu\leqslant\int\Big(f_{m+1}-\sum_{i=1}^{m}k_if_i\Big)\phi^*d\mu,$$

其中 ϕ^* 为对应于点 $(c_1,\cdots,c_m,c^*)^\tau$ 的检验函数. 这蕴含了当 $f_{m+1}-\sum_{i=1}^{m}k_if_i$ 为正时 $\phi^*=1$, 而当其为负时 $\phi^*=0$, 即 ϕ^* 满足 (3.12) a.e. μ.

(v) 是与 (ii) 对偶的命题, 无需再加以证明. ■

3.4 无偏检验

3.4.1 检验的无偏性

在本节中, 我们讨论更广泛的一类检验, 即无偏检验的情形.

定义 3.3 称 ϕ 是检验问题 (3.1) 的一个水平为 α 的**无偏检验**, 如果其功效函数 $\beta_\phi(\theta)=\mathrm{E}_\theta[\phi(X)]$ 满足条件

$$\begin{aligned}&\beta_\phi(\theta)\leqslant\alpha,\ \text{若}\ \theta\in\Theta_0,\\&\beta_\phi(\theta)\geqslant\alpha,\ \text{若}\ \theta\in\Theta_1.\end{aligned}\tag{3.14}$$

亦即满足条件

$$\sup_{\theta\in\Theta_0}\beta_\phi(\theta)\leqslant\alpha\leqslant\inf_{\theta\in\Theta_1}\beta_\phi(\theta).$$

简言之, 一个合理的检验在原假设成立时它被否定的概率, 不应超过原假设不成立时, 原假设被否定的概率.

现在换一个角度来看检验的 "无偏性". 为此, 设 $X\sim N(\theta,1),\ -\infty<\theta<\infty$, 考虑检验问题 $H:\theta=0\leftrightarrow K:\theta\neq0$. 易见以 $\{X>u(\alpha)\}$ 为否定域的检验 ϕ_1 和以 $\{X<-u(\alpha)\}$ 为否定域的检验 ϕ_2 都有水平 α, 前者对 K 的右边部分 $\{\theta>0\}$ 是 UMP 的, 很有效, 但在 $\{\theta<0\}$ 上的功效却很差, 不仅小于 α, 且随 $\theta\to-\infty$ 而趋于 0, 即它过分偏重于 K 的 $\{\theta>0\}$ 的部分而忽视了另一部分. 检验 ϕ_2 则相反. 如果采取折中, 即取 $\{|X|>u(\alpha/2)\}$ 为否定域, 则我们对这两部分都没有偏向, 是为无偏. 为此付出的代价是所采用的检验在 K 的两个部分均未达到可能的最大功效.

为了下面的应用, 我们引进相似性与 Neyman 结构两个概念.

设 Θ 为欧氏的, Θ_0 和 Θ_1 有公共边界 ω, ϕ 是检验问题 (3.1) 的一个水平为 α 的无偏检验, 其功效函数 $\beta_\phi(\theta)$ 连续, 则

$$\mathrm{E}_\theta[\phi(X)]\equiv\alpha,\quad\theta\in\omega.\tag{3.15}$$

满足这一条件的检验 ϕ 称为在 H 和 K 的边界 ω 上是相似的.

一般地, 设 ω 为参数空间 Θ 的一个子集, ϕ 为一个检验, 其功效函数在 ω 上恒为常数 α, 则称 ϕ **对于集合** ω **相似**, 更确切地说, 是**对分布族** $\{P_\theta,\theta\in\omega\}$ **相似**.

今设 $\{P_\theta,\theta\in\omega\}$ 有一充分统计量 $T=T(X)$, 则由习题 1.23, 可选取 $\mathrm{E}_\theta[\phi(X)|t]$ 的一个版本 $\mathrm{E}[\phi(X)|t]$ 与 θ 无关, 如果

$$\mathrm{E}[\phi(X)|t]=\alpha\ (\text{常数}),\ \text{a.s. } P_\theta,\theta\in\omega,\tag{3.16}$$

则 (3.15) 成立, 因而检验 ϕ 对于集合 ω 相似. 称满足 (3.16) 的检验 ϕ **在** ω **上关于** T **有 Neyman 结构**, 简称**对** (T,ω) **有 Neyman 结构**.

相似性与 Neyman 结构有以下联系.

定理 3.6 ϕ 对 (T,ω) 有 Neyman 结构 $\Rightarrow$ ϕ 对于集合 ω 相似; 反之, 若 T 对 $\{P_\theta,\theta\in\omega\}$ 除具充分性外还是有界完全的, 则相反的蕴含关系成立.

证明 前一结论显然. 现设 ϕ 对于 ω 相似, $\mathrm{E}_\theta[\phi(X)]=\alpha$, $\theta\in\omega$, 且设 T 满足定理条件. 如前所述, 当 $\theta\in\omega$ 时, 存在 $\mathrm{E}_\theta[\phi(X)|t]$ 的一个版本 $\psi(T):=\mathrm{E}[\phi(X)|t]$ 与 θ 无关. 令 $h(T)=\psi(T)-\alpha$, 则由 ϕ 的相似性, $\mathrm{E}_\theta[h(T)]=\mathrm{E}_\theta\{\mathrm{E}[\phi(X)|T]\}-\alpha=0,\theta\in\omega$. 由 T 的有界完全性, $h(T)=0$, a.s. $P_\theta,\theta\in\omega$, 即 (3.16) 成立. 证毕. ■

定义 3.4 设 ϕ 是检验问题 (3.1) 的一个水平为 α 的无偏检验, 且对 (3.1) 的任一水平为 α 的无偏检验 ψ 有

$$\beta_\psi(\theta)\leqslant\beta_\phi(\theta),\ \theta\in\Theta_1,$$

则称 ϕ 为检验问题 (3.1) 的一个水平为 α 的**一致最优无偏检验** (Uniformly Most Powerful Unbiased Test), 简称 **UMP 无偏检验**或 **UMPU 检验**.

易见, 水平为 α 的 UMP 检验必为水平为 α 的 UMP 无偏检验. 有一大类检验问题, 其中 UMP 检验不存在, 但存在 UMP 无偏检验. 特别是在检验某些形如 $\theta\leqslant\theta_0$ 或 $\theta=\theta_0$ 的假设, 而又涉及其他参数时更是这样.

3.4.2 单参数指数族的 UMP 无偏检验

考虑单参数指数族 $dP_\theta(x)=C(\theta)\exp\{\theta T(x)\}d\mu(x),\theta\in\Theta$, 其中自然参数空间 Θ 为 $\mathbf{R}$ 中的一个非退化区间. 设 $\theta_0,\theta_1,\theta_2$ 均为该区间中给定的点, 且 $\theta_1<\theta_2$. 如前所示, 在此模型中, 单边假设有 UMP 检验. 不难证明, 双边假设

$$H_1:\theta_1\leqslant\theta\leqslant\theta_2\leftrightarrow K_1:\theta\notin[\theta_1,\theta_2],$$
$$H_2:\theta=\theta_0\leftrightarrow K_2:\theta\neq\theta_0$$

都没有 UMP 检验. 下面将证明, 它们都有 UMP 无偏检验. 为此, 先引进以下引理.

引理 3.7 设对于分布族 $\{P_\theta\}$, 检验问题 $H\leftrightarrow K$ 的每个检验的功效函数都是连续的, 若 ϕ_0 是水平为 α 的, 且在满足 (3.15) 的所有检验中是 UMP 的, 则 ϕ_0 是 UMP 无偏的.

证明 与恒等于 α 的检验比较, ϕ_0 的功效不小于 α, 故 ϕ_0 是水平 α 的无偏检验. 而满足 (3.15) 的检验的类包括水平 α 的所有无偏检验, 故 ϕ_0 是 UMP 无偏的. ■

定理 3.8 设检验 ϕ 形如

$$\phi(x)=\begin{cases}1, & \text{当 } T(x)\notin[c_1,c_2],\\ \gamma_i, & \text{当 } T(x)=c_i,i=1,2,\\ 0, & \text{当 } c_1<T(x)<c_2\end{cases}\tag{3.17}$$

且满足

$$\mathrm{E}_{\theta_i}[\phi(X)]=\alpha\ \Big(\text{即}\int\phi(x)C(\theta_i)\exp\{\theta_iT(x)\}d\mu(x)=\alpha\Big),\ i=1,2,\tag{3.18}$$

其中 $c_1<c_2,0\leqslant\gamma_i\leqslant1,i=1,2$, 则 ϕ 是检验问题 $H_1\leftrightarrow K_1$ 的水平为 α 的 UMP 无偏检验.

证明 记 $t=T(x)$. 取定 $\vartheta\notin[\theta_1,\theta_2]$. 往证: 存在常数 k_1,k_2 (可与 ϑ 有关), 使

$$\phi(x)=\begin{cases}1, & \text{当 } C(\vartheta)\mathrm{e}^{\vartheta t}>k_1C(\theta_1)\mathrm{e}^{\theta_1 t}+k_2C(\theta_2)\mathrm{e}^{\theta_2 t},\\ 0, & \text{当 } C(\vartheta)\mathrm{e}^{\vartheta t}<k_1C(\theta_1)\mathrm{e}^{\theta_1 t}+k_2C(\theta_2)\mathrm{e}^{\theta_2 t}.\end{cases}\tag{3.19}$$

为确定起见设 $\vartheta>\theta_2$. 定义函数

$$H(t)=C(\vartheta)\exp\{(\vartheta-\theta_2)t\}-k_1C(\theta_1)\exp\{(\theta_1-\theta_2)t\}-k_2C(\theta_2),$$

其中 k_1,k_2 由方程组 $H(c_1)=H(c_2)=0$ 确定. 容易验证, 该方程组有解且 $k_1<0$, 故函数 H 在 $(-\infty,\infty)$ 上严格凸, 由此可知

$$\begin{aligned}&t\notin[c_1,c_2]\text{ 或 }c_1<t<c_2,\text{当且仅当 }H(t)>0\text{ 或 }H(t)<0,\\ &\text{当且仅当 }C(\vartheta)\mathrm{e}^{\vartheta t}>k_1C(\theta_1)\mathrm{e}^{\theta_1 t}+k_2C(\theta_2)\mathrm{e}^{\theta_2 t}\\ &\text{或 }C(\vartheta)\mathrm{e}^{\vartheta t}<k_1C(\theta_1)\mathrm{e}^{\theta_1 t}+k_2C(\theta_2)\mathrm{e}^{\theta_2 t},\end{aligned}$$

对照 (3.17) 即得 (3.19).

由此及定理 3.5, 满足 (3.18) 和 (3.19) 的 ϕ, 在满足 (3.18) 的检验函数类中, 使 $\beta_\phi(\vartheta),\vartheta\notin[\theta_1,\theta_2]$ 一致地达到最大. 由指数族的性质, 任一检验的功效函数连续. 故 ϕ 如有水平 α, 则由引理 3.7, 必为水平 α 的 UMP 无偏检验. 因此, 为证本定理, 只要证明, ϕ 有水平 α.

为此, 取定 $\vartheta\in(\theta_1,\theta_2)$. 一个对偶的论证即得, 在满足 (3.18) 的检验函数类中, 以满足 (3.19) 的那一个使 $\beta_\phi(\vartheta)$ 达到最小, 且不超过 α, 即 ϕ 有水平 α. 证毕. ■

现在考虑 $H_2\leftrightarrow K_2$, 给定 $\alpha\in(0,1)$, $\theta_0\in\Theta$.

定理 3.9 *设检验 ϕ 形如 (3.17) 且满足*

$$\mathrm{E}_{\theta_0}[\phi(X)]=\alpha,\quad \mathrm{E}_{\theta_0}[T(X)\phi(X)]=\alpha\mathrm{E}_{\theta_0}[T(X)],\tag{3.20}$$

则 ϕ 是 $H_2\leftrightarrow K_2$ 的水平 α 的 UMP 无偏检验.

证明 水平为 α 的无偏检验 ϕ 必满足 (3.20), 其中前一式由功效函数的连续性得出, 后一式是因其功效函数在 θ_0 取到最小值, 注意到 $\int_{\mathcal{X}}C(\theta)\exp\{\theta T(x)\}\phi(x)d\mu$ 可在积分号下对 θ 求导, 且 $C'(\theta)/C(\theta)=-\mathrm{E}_\theta[T(X)]$ 即得. 简记 $t=T(x)$, (3.20) 即为 $\int\phi(x)C(\theta_0)\mathrm{e}^{\theta_0 t}d\mu=\alpha$, 且 $\int\phi(x)C(\theta_0)t\mathrm{e}^{\theta_0 t}d\mu=\alpha\mathrm{E}_{\theta_0}[T(X)]$.

任取 $\vartheta\neq\theta_0$. 与前一定理的证明类似, 只需证明存在常数 k_1,k_2, 使

$$\phi(x)=\begin{cases}1, & \text{当 } C(\vartheta)\mathrm{e}^{\vartheta t}>(k_1+k_2t)C(\theta_0)\mathrm{e}^{\theta_0 t},\\ 0, & \text{当 } C(\vartheta)\mathrm{e}^{\vartheta t}<(k_1+k_2t)C(\theta_0)\mathrm{e}^{\theta_0 t}.\end{cases}\tag{3.21}$$

再用定理 3.5 和引理 3.7 即得所要的结论.

记 $H(t)=C(\vartheta)\exp\{(\vartheta-\theta_0)t\}-C(\theta_0)(k_1+k_2t)$, 其中 k_1,k_2 由 $H(c_1)=H(c_2)=0$ 确定. 因 $c_1<c_2$, 容易验证由该方程组可解出 k_1,k_2. 由 H 在 $(-\infty,\infty)$ 上为严格凸, $H(c_1)=H(c_2)=0$, 立得

$$\begin{aligned}&t\notin[c_1,c_2]\text{ 或 }c_1<t<c_2,\text{当且仅当 }H(t)>0\text{ 或 }H(t)<0,\\ &\text{当且仅当 }C(\vartheta)\mathrm{e}^{\vartheta t}>(k_1+k_2t)C(\theta_0)\mathrm{e}^{\theta_0 t}\text{ 或 }C(\vartheta)\mathrm{e}^{\vartheta t}<(k_1+k_2t)C(\theta_0)\mathrm{e}^{\theta_0 t},\end{aligned}$$

对照 (3.17) 即得 (3.21). 证毕. ■

注 3.1 由定理 3.9 的证明可以看出, 在单参数指数族中, 满足 (3.21) 的一个检验函数 ϕ, 只要是无偏的且水平为 α, 则必为 UMP 无偏的. 对定理 3.8 也有类似的诠释. 此时 (3.20) 可代之以如下的关系式:

$$\beta_\phi(\theta_0)=\alpha,\quad d\beta_\phi(\theta)/d\theta|_{\theta=\theta_0}=0.$$

例 3.4 下面所说的接受域或检验函数, 分别是指有关检验问题的水平为 α 的 UMP 无偏检验的接受域或检验函数.

1° 设 $X_1,\cdots,X_n$ 为取自正态总体 $\{N(\theta,\sigma^2),\theta\in\mathbf{R}\}$ 的 i.i.d. 样本, σ^2 已知.

$H_1\leftrightarrow K_1$: 接受域可表示为 $\{c_1\leqslant\bar X\leqslant c_2\}$, c_1,c_2 满足:

$$\Phi[\sqrt{n}(c_2-\theta_i)/\sigma]-\Phi[\sqrt{n}(c_1-\theta_i)/\sigma]=1-\alpha,\quad i=1,2.$$

此处 $\Phi(\cdot)$ 为标准正态分布函数.

$H_2\leftrightarrow K_2$: 接受域可表示为 $\{|\bar X-\theta_0|\leqslant u(\alpha/2)\sigma/\sqrt{n}\}$.

2° 设 $X_1,\cdots,X_n$ 为取自指数分布总体 $\{E(0,\theta),\theta>0\}$ 的 i.i.d. 样本. 在 P_θ 之下, $2\sum_{i=1}^n X_i/\theta\sim\chi^2_{2n}$, 其分布函数和密度函数分别记为 $K_{2n}(\cdot)$ 和 $k_{2n}(\cdot)$.

$H_1\leftrightarrow K_1$: 接受域可表示为 $\{c_1\leqslant\sum_{i=1}^n X_i\leqslant c_2\}$, c_1,c_2 满足:

$$K_{2n}(2c_2/\theta_i)-K_{2n}(2c_1/\theta_i)=1-\alpha,\quad i=1,2.$$

$H_2\leftrightarrow K_2$: 接受域如上所示, 但 c_1,c_2 满足:

$$\begin{aligned}&K_{2n}(2c_2/\theta_0)-K_{2n}(2c_1/\theta_0)=1-\alpha,\\&\int_{2c_1/\theta_0}^{2c_2/\theta_0}tk_{2n}(t)dt=2n(1-\alpha).\end{aligned}$$

其中第二个方程也可代之以 $c_1k_{2n}(2c_1/\theta_0)=c_2k_{2n}(2c_2/\theta_0)$.

3° 设样本 $X\sim B(n,\theta),0\leqslant\theta\leqslant1$. 检验函数 $\phi(x)$ 形如 (3.17), 其中 $T(X)=X$, 而 c_i, γ_i 由

$$\sum_{t=c_1+1}^{c_2-1}\binom{n}{t}\theta_i^t(1-\theta_i)^{n-t}+\sum_{j=1}^{2}(1-\gamma_j)\binom{n}{c_j}\theta_i^{c_j}(1-\theta_i)^{n-c_j}=1-\alpha,\quad i=1,2$$

确定 (对 $H_1\leftrightarrow K_1$), 或由

$$\begin{aligned}&\sum_{t=c_1+1}^{c_2-1}\binom{n}{t}\theta_0^t(1-\theta_0)^{n-t}+\sum_{j=1}^{2}(1-\gamma_j)\binom{n}{c_j}\theta_0^{c_j}(1-\theta_0)^{n-c_j}=1-\alpha,\\&\sum_{t=c_1+1}^{c_2-1}t\binom{n}{t}\theta_0^t(1-\theta_0)^{n-t}+\sum_{j=1}^{2}(1-\gamma_j)c_j\binom{n}{c_j}\theta_0^{c_j}(1-\theta_0)^{n-c_j}=n\theta_0(1-\alpha)\end{aligned}$$

确定 (对 $H_2\leftrightarrow K_2$).

3.4.3 多参数指数族的 UMP 无偏检验

设样本 X 有分布

$$dP^X_{\theta,\vartheta}(x)=C(\theta,\vartheta)\exp\{\theta U(x)+\vartheta^\tau V(x)\}d\mu(x),\quad(\theta,\vartheta^\tau)^\tau\in\Theta,\tag{3.22}$$

此处 μ 为样本空间 $(\mathcal{X},\mathcal{B})$ 上的一个 σ 有限测度, ϑ 的维数为 $d\geqslant 1$, Θ 是自然参数空间, 在 $\mathbf{R}^{d+1}$ 中的内点集 Θ^o 非空. (U,V) 为完全充分统计量. 换到 (U,V) 的样本空间 $(\mathbf{R}^{d+1},\mathcal{B}_{d+1})$, 则

$$(U,V)\sim dP_{\theta,\vartheta}(u,v)=C(\theta,\vartheta)\exp\{\theta u+\vartheta^\tau v\}d\nu(u,v),(\theta,\vartheta)\in\Theta.$$

由引理 1.14, 有

$$V\sim dP^V_{\theta,\vartheta}(v)=C_\theta(\vartheta)\exp(\vartheta^\tau v)d\nu_\theta(v), \tag{3.23}$$

给定 $V=v$ 时 U 的条件分布为

$$U|v\sim dP^{U|v}_\theta(u)=C_v(\theta)\mathrm{e}^{\theta u}d\nu_v(u), \tag{3.24}$$

即它是一个参数为 θ 的单参数指数族, 与 ϑ 无关.

本节将就这一分布族讨论关于参数 θ 的四个检验问题:

$$H_1:\theta\leqslant\theta_0\leftrightarrow K_1:\theta>\theta_0,$$

$$H_2:\theta\leqslant\theta_1\text{ 或 }\theta\geqslant\theta_2\leftrightarrow K_2:\theta_1<\theta<\theta_2,$$

$$H_3:\theta_1\leqslant\theta\leqslant\theta_2\leftrightarrow K_3:\theta\notin[\theta_1,\theta_2].$$

$$H_4:\theta=\theta_0\leftrightarrow K_4:\theta\neq\theta_0,$$

此处, $\theta_0,\theta_1,\theta_2$ 均为给定的点, 且 $\theta_1<\theta_2$; 而 ϑ 则称为多余参数, 它的出现往往会使关于参数 θ 的统计推断复杂化.

在这类问题的统计分析中, $U|v$ 的上述条件分布起着关键的作用, 据此可通过条件检验将问题化为已知情形. 水平为 α 的 UMP 无偏检验分别如下.

(i) 对 $H_1\leftrightarrow K_1$:

$$\phi_1(u,v)=\begin{cases}1, & \text{当 } u>c(v),\\ \gamma(v), & \text{当 } u=c(v),\\ 0, & \text{当 } u<c(v).\end{cases}$$

其中 $c(v),0\leqslant\gamma(v)\leqslant 1$ 由 $\mathrm{E}_{\theta_0}[\phi_1(U,V)|v]=\alpha$ 确定, 条件分布如 (3.24) 所示.

(ii) 对 $H_2\leftrightarrow K_2$:

$$\phi_2(u,v)=\begin{cases}1, & \text{当 } c_1(v)<u<c_2(v),\\ \gamma_i(v), & \text{当 } u=c_i(v),i=1,2,\\ 0, & \text{当 } u\notin[c_1(v),c_2(v)],\end{cases}$$

其中 $c_i(v),\gamma_i(v)\in[0,1]$ $(i=1,2)$ 由 $\mathrm{E}_{\theta_1}[\phi_2(U,V)|v]=\mathrm{E}_{\theta_2}[\phi_2(U,V)|v]=\alpha$ 确定.

(iii) 对 $H_3\leftrightarrow K_3$:

$$\phi_3(u,v)=\begin{cases}1, & \text{当 } u\notin[c_1(v),c_2(v)],\\ \gamma_i(v), & \text{当 } u=c_i(v),i=1,2,\\ 0, & \text{当 } c_1(v)<u<c_2(v),\end{cases}$$

其中 $c_i(v),\gamma_i(v)\in[0,1]$ $(i=1,2)$ 由 $\mathrm{E}_{\theta_1}[\phi_3(U,V)|v]=\mathrm{E}_{\theta_2}[\phi_3(U,V)|v]=\alpha$ 确定.

(iv) 对 $H_4\leftrightarrow K_4$: $\phi_4(u,v)$ 仍有 $\phi_3(u,v)$ 的形式, 但 $c_i(v),\gamma_i(v)$ 由 $\mathrm{E}_{\theta_0}[\phi_4(U,V)|v]=\alpha$ 及 $\mathrm{E}_{\theta_0}[U\phi_4(U,V)|v]=\alpha\mathrm{E}_{\theta_0}(U|v)$ 确定.

定理 3.10 设 Θ 在 $\mathbf{R}^{d+1}$ 中的内点集 Θ^{o} 非空, 且在假设中出现 θ_j 的问题中, $\mathbf{R}^{d+1}$ 中相应的超平面 $\theta=\theta_j$ 与 Θ^{o} 均有非空的交, $j\in\{0,1,2\}$, 则上述检验 ϕ_i 为 $H_i\leftrightarrow K_i$ 的水平为 α 的 UMP 无偏检验, $i=1,2,3,4$.

证明 考虑 $H_1\leftrightarrow K_1$, 记 $\omega_0=\{\vartheta:(\theta_0,\vartheta^\tau)^\tau\in\Theta\}$. 由假定, 它在 $\mathbf{R}^d$ 中有内点. 由 (3.23), 若固定 $\theta=\theta_0$, 则 V 在参数空间 ω_0 上的分布族为

$$\mathcal{P}_0^V=\{C_{\theta_0}(\vartheta)\exp(\vartheta^\tau v)d\nu_{\theta_0}(v),\vartheta\in\omega_0\},$$

V 为其完全充分统计量.

现设 $\phi(u,v)$ 是 $H_1\leftrightarrow K_1$ 的任一水平为 α 的无偏检验, 则因 ϕ 的功效函数 $\mathrm{E}_{\theta,\vartheta}\phi(U,V)$ 连续, 故它在 ω_0 上取常数值 α. 由此及定理 3.6,

$$\mathrm{E}_{\theta_0}[\phi(U,V)|v]\equiv\mathrm{E}_{\theta_0,\vartheta}[\phi(U,V)|v]=\alpha,\quad \text{a.e. }\mathcal{P}_0^V, \tag{3.25}$$

因 (3.24) 中的分布族 $dP_\theta^{U|v}$ 与 ϑ 无关, (3.25) 只有一个公共的例外集 B. 再由指数族的性质, 以及 B 为 $\mathcal{P}_0^V$-零概集, 推出对所有 (θ,ϑ), $P_{\theta,\vartheta}(B)=0$. 故不妨设 B 为空集.

因给定 v 时 U 的条件分布为只依赖于单参数 θ 的指数族, 故由定理 3.4 及 (3.25), 有

$$\mathrm{E}_{\theta,\vartheta}[\phi_1(U,V)|v]\geqslant\mathrm{E}_{\theta,\vartheta}[\phi(U,V)|v],\quad \theta>\theta_0,\ (\theta,\vartheta)\in\Theta,$$

$$\mathrm{E}_{\theta,\vartheta}[\phi_1(U,V)|v]\leqslant\alpha,\quad \theta\leqslant\theta_0,(\theta,\vartheta)\in\Theta.$$

故对任一水平为 α 的无偏检验 ϕ, 当 $(\theta,\vartheta)\in\Theta$ 时有

$$\mathrm{E}_{\theta,\vartheta}[\phi_1(U,V)]\leqslant\alpha,\quad \theta\leqslant\theta_0,$$

$$\mathrm{E}_{\theta,\vartheta}[\phi_1(U,V)]\geqslant\mathrm{E}_{\theta,\vartheta}[\phi(U,V)],\quad \theta>\theta_0.$$

故 ϕ_1 即为 $H_1\leftrightarrow K_1$ 的水平为 α 的 UMP 无偏检验.

$H_i\leftrightarrow K_i$ $(i=2,3,4)$ 的论证与此完全相似, 从略. ■

注 3.2 以上的理论只能处理单参数的检验问题, 即使对指数族而言, 多参数检验问题的 UMP 无偏检验一般也多不存在. 由于经参数的线性变换, 指数族仍变为指数族, 故以上理论适用于参数的线性组合, 即 $\theta^*=a_0\theta+\sum_{i=1}^d a_i\vartheta_i$, 其中 $a_0,\cdots,a_d$ 不全为 0. 例如, 设 $a_0\neq0$, 则容易把 (3.22) 写成

$$dP_{\theta^*,\vartheta}^X(x)=C(\theta^*,\vartheta)\exp\{\theta^*U^*(x)+\vartheta^\tau V^*(x)\}d\mu(x),$$

的形式, 即可纳入已经建立的理论框架.

在涉及无偏检验的一些应用中, 在一般的指数族模型 (3.22) 中, 常常存在一个统计量 $W=h(U,V)$, 当 $\theta=\theta_0$ 时与 V 独立, 且对每个给定的 v, 在检验假设 $H_1:\theta\leqslant\theta_0$ 的场合, $w=h(u,v)$ 为 u 的增函数; 而在检验假设 $H_4:\theta=\theta_0$ 的场合, $w=h(u,v)$ 对 u 为线性, 则水平为 α 的 UMP 无偏检验分别满足下列条件.

(i) 对 $H_1 \leftrightarrow K_1$:

$$\phi_1(w) = \begin{cases} 1, & \text{当 } w > c, \\ \gamma, & \text{当 } w = c, \\ 0, & \text{当 } w < c. \end{cases}$$

其中 c, γ 由 $\mathrm{E}_{\theta_0}[\phi_1(W)] = \alpha$ 确定.

(iv) 对 $H_4 \leftrightarrow K_4$:

$$\phi_4(w) = \begin{cases} 1, & \text{当 } w \notin [c_1, c_2], \\ \gamma_i, & \text{当 } w = c_i, i = 1, 2, \\ 0, & \text{当 } c_1 < w < c_2, \end{cases}$$

其中 c_i, γ_i $(i = 1, 2)$ 由 $\mathrm{E}_{\theta_0}[\phi_4(W)] = \alpha$ 及 $\mathrm{E}_{\theta_0}[W\phi_4(W)] = \alpha\mathrm{E}_{\theta_0}(W)$ 确定.

$H_2 : \theta \leqslant \theta_1$ 或 $\theta \geqslant \theta_2$ 和 $H_3 : \theta_1 \leqslant \theta \leqslant \theta_2$ 的化简需要 $w = h(u, v)$ 对任意给定的 v 关于 u 单调增, 且当 $\theta = \theta_1$ 和 $\theta = \theta_2$ 时 W 与 V 独立. 此时水平为 α 的 UMP 无偏检验分别满足下列条件.

(iii) 对 $H_3 \leftrightarrow K_3$: ϕ_3 与 ϕ_4 有相同的形式, 但 c_i, γ_i $(i = 1, 2)$ 由 $\mathrm{E}_{\theta_1}[\phi_4(W)] = \mathrm{E}_{\theta_2}[\phi_4(W)] = \alpha$ 确定.

(ii) 对 $H_2 \leftrightarrow K_2$: $\phi_2(w, \alpha) = 1 - \phi_3(w, 1 - \alpha)$.

参照 Lehmann (1986) 定理 5.1, 在上述情况下, 我们有下面的定理.

定理 3.11 *设 X 的分布如 (3.22) 所示, 且定理 3.10 的条件成立. 若当 $\theta = \theta_0$ 时 $W = h(U, V)$ 与 V 独立, 则 ϕ_1 对 $H_1 \leftrightarrow K_1$ 是 UMP 无偏的, 只要对每个 v, 函数 h 对 u 是增的; 而 ϕ_4 对 $H_4 \leftrightarrow K_4$ 是 UMP 无偏的, 只要*

$$h(u, v) = a(v)u + b(v), \quad a(v) > 0.$$

检验 ϕ_2 和 ϕ_3 对 $H_2 \leftrightarrow K_2$ 和 $H_3 \leftrightarrow K_3$ 是 UMP 无偏的, 如果当 $\theta = \theta_1$ 和 $\theta = \theta_2$ 时 W 与 V 独立, 且对每个 v, 函数 h 对 u 是增的.

证明 根据定理的条件, 这些检验分别等价于定理 3.10 中相应的检验, 以检验问题 $H_4 \leftrightarrow K_4$ 为例, 涉及的常数 $c_i(v), \gamma_i(v)$ 由 $\mathrm{E}_{\theta_0}[\phi_4(W, v)|v] = \alpha$ 和

$$\mathrm{E}_{\theta_0}\Big[\phi_4(W, v)\frac{W - b(v)}{a(v)}\Big|v\Big] = \alpha\mathrm{E}_{\theta_0}\Big[\frac{W - b(v)}{a(v)}\Big|v\Big]$$

确定, 后者可简化为 $\mathrm{E}_{\theta_0}[W\phi_4(W, v)|v] = \alpha\mathrm{E}_{\theta_0}[W|v]$. 因为当 $\theta = \theta_0$时 W 与 V 独立, 故这些条件期望都简化为无条件期望, 而 $c_i(v), \gamma_i(v)$ 都是与 v 无关的常数. ■

3.4.4 与正态有关的检验

3.4.3 节我们讨论了在指数族 (3.22), 即

$$dP_{\theta,\vartheta}(x) = C(\theta, \vartheta)\exp\{\theta U(x) + \vartheta^\tau V(x)\}d\mu(x), \quad (\theta, \vartheta^\tau)^\tau \in \Theta$$

中关于 θ 的检验问题 $H_i \leftrightarrow K_i$ $(i = 1, \cdots, 4)$ 的 UMP 无偏检验的一般理论, 其中 $\Theta \subset \mathbf{R}^{d+1}$ 为自然参数空间, 满足定理 3.10 的条件. 本节要在此基础上讨论正态分布参数的检验问题.

1. 一样本 t-检验

设 $X_1,\cdots,X_n$ 为取自正态总体 $N(a,\sigma^2)$ 的 i.i.d. 样本, σ^2 未知, 要检验假设 $a\leqslant 0\leftrightarrow a>0$.

$(X_1,\cdots,X_n)^\tau$ 的分布关于 n 维 Lebesgue 测度的密度

$$(2\pi\sigma^2)^{-n/2}\exp\Big[\frac{-na^2}{2\sigma^2}\Big]\exp\Big[-\frac{1}{2\sigma^2}\sum_{i=1}^n x_i^2+\frac{a}{\sigma^2}\sum_{i=1}^n x_i\Big],$$

构成一个形如 (3.22) 的两参数指数族, 其中

$$U=\sum_{i=1}^n X_i,\quad V=\sum_{i=1}^n X_i^2,\quad \theta=a/\sigma^2,\quad \vartheta=-1/(2\sigma^2).$$

上述假设转化为 $\theta\leqslant 0\leftrightarrow\theta>0$. 由定理 3.10, 水平为 α 的 UMP 无偏检验有接受域 $U\leqslant c_1(V)$. 令 $S^2=\sum_{i=1}^n(X_i-\bar X)^2/(n-1)=(V-U^2/n)/(n-1)$, $T=\sqrt{n}\bar X/S=\sqrt{n-1}U/\sqrt{n(V-U^2/n)}$. 因 $u/\sqrt{v-u^2/n}$ 为 u 的增函数, 上述接受域等价于 $T\leqslant c(V)$, $c(\cdot)$ 由 $a=0$ 时 (T,V) 的分布族确定. 因当 $a=0$ 时, T 的分布与 σ^2 无关, 故由 Basu 定理, 此时 T 与 V 独立, 且 $T\sim t_{n-1}$. 故 $c(v)$ 应取常数且为 $t_{n-1}(\alpha)$, 即分布 t_{n-1} 的上 α 分位数. 因而检验问题 $a\leqslant 0\leftrightarrow a>0$ 的水平为 α 的 UMP 无偏检验的接受域为 $T\leqslant t_{n-1}(\alpha)$.

由类似的论证可得, 对于双边假设 $a=0\leftrightarrow a\neq 0$, 水平为 α 的 UMP 无偏检验的接受域为 $|T|\leqslant t_{n-1}(\alpha/2)$.

2. 两样本 t-检验

设 $X_1,\cdots,X_m$ 和 $Y_1,\cdots,Y_n$ 分别为取自正态总体 $N(a,\sigma^2)$ 和 $N(b,\sigma^2)$ 的 i.i.d. 样本, σ^2 未知, 且两样本独立, 要检验假设 $a\leqslant b\leftrightarrow a>b$ 和 $a=b\leftrightarrow a\neq b$.

令 $S^2=[\sum_{i=1}^m(X_i-\bar X)^2+\sum_{i=1}^n(Y_j-\bar Y)^2]/(m+n-2)$. 应用定理 3.10, 与一样本情形类似, 可得上述两个检验问题的水平为 α 的 UMP 无偏检验的接受域分别为

$$\sqrt{mn/(m+n)}(\bar X-\bar Y)/S\leqslant t_{m+n-2}(\alpha)$$

和

$$|\sqrt{mn/(m+n)}(\bar X-\bar Y)/S|\leqslant t_{m+n-2}(\alpha/2).$$

3. 一样本方差检验 (χ^2-检验)

设 $X_1,\cdots,X_n$ 为取自正态总体 $N(a,\sigma^2)$ 的 i.i.d. 样本, a 和 σ^2 均未知, 要检验假设 $\sigma^2\leqslant\sigma_0^2\leftrightarrow\sigma^2>\sigma_0^2$, $\sigma^2\geqslant\sigma_0^2\leftrightarrow\sigma^2<\sigma_0^2$ 和 $\sigma^2=\sigma_0^2\leftrightarrow\sigma^2\neq\sigma_0^2$. 读者不难写出这三个问题的水平为 α 的 UMP 无偏检验.

以第三个问题为例, 由定理 3. 10, 考虑给定 $\bar X$ 条件下基于 $\sum_{i=1}^n(X_i-\bar X)^2$ 的条件检验, 知所求检验具有形如

$$c_1(\bar X)\sigma_0^2\leqslant\sum_{i=1}^n(X_i-\bar X)^2\leqslant c_2(\bar X)\sigma_0^2$$

的接受域. 由于 $\bar X$ 与 $\sum_{i=1}^n(X_i-\bar X)^2$ 独立, $c_1(\bar X)\equiv c_1, c_2(\bar X)\equiv c_2$ 为常数且由关

系式

$$K_{n-1}(c_2)-K_{n-1}(c_1)=1-\alpha,\quad c_1k_{n-1}(c_1)=c_2k_{n-1}(c_2)$$

决定. 其中 K_{n-1} 和 k_{n-1} 分别为 χ^2_{n-1} 变量的分布函数和密度函数. 上面第二式是缘于无偏性, 要求在 $\sigma=\sigma_0$ 处 $\mathrm{E}_\sigma[\phi(X)]$ 取得最小值, 即在该点

$$P_\sigma\{c_1\sigma_0^2/\sigma^2\leqslant\sum_{i=1}^n(X_i-\bar X)^2/\sigma^2\leqslant c_2\sigma_0^2/\sigma^2\}$$

取得最大值. 记 $r=\sigma_0^2/\sigma^2$, 注意到在 P_σ 之下 $\sum_{i=1}^n(X_i-\bar X)^2/\sigma^2\ \sim\chi^2_{n-1}$, 则由极值点的性质,

$$\frac{d}{dr}\int_{rc_1}^{rc_2}k_{n-1}(t)dt\bigg|_{r=1}=0,$$

由此即得上面第二式 (参看注 3.1).

在应用中, 一般使用更简单的取法: $c_1=\chi^2_{n-1}(1-\alpha/2), c_2=\chi^2_{n-1}(\alpha/2)$. 所得的检验虽不是无偏的, 但与上面求得的最优无偏检验相差不大.

类似地, 假设 $\sigma^2\leqslant\sigma_0^2\leftrightarrow\sigma^2>\sigma_0^2$ 的水平为 α 的 UMP 无偏检验的接受域为:

$$\sum_{i=1}^n(X_i-\bar X)^2\leqslant\chi^2_{n-1}(\alpha)\sigma_0^2.$$

4. 两样本方差检验 (F-检验)

设 $X_1,\cdots,X_m$ 和 $Y_1,\cdots,Y_n$ 分别为取自正态总体 $N(a,\sigma_1^2)$ 和 $N(b,\sigma_2^2)$ 的 i.i.d. 样本, $a,b,\sigma_1^2>0,\sigma_2^2>0$ 未知, 且两样本独立,

要检验假设 $\sigma_1^2\leqslant\sigma_2^2\leftrightarrow\sigma_1^2>\sigma_2^2$ 和 $\sigma_1^2=\sigma_2^2\leftrightarrow\sigma_1^2\neq\sigma_2^2$.

$(X_1,\cdots,X_m,Y_1,\cdots,Y_n)$ 的联合密度为

$$C(a,b,\sigma_1,\sigma_2)\exp\Big[-\frac{1}{2\sigma_1^2}\sum_{i=1}^m x_i^2-\frac{1}{2\sigma_2^2}\sum_{j=1}^n y_j^2+\frac{ma}{\sigma_1^2}\bar x+\frac{nb}{\sigma_2^2}\bar y\Big],$$

它可写成注 3.2 中的那种形式, 其中

$$\theta^*=\frac{1}{2\sigma_1^2}-\frac{1}{2\sigma_2^2},\quad \vartheta=\Big(-\frac{1}{2\sigma_1^2},\frac{ma}{\sigma_1^2},\frac{nb}{\sigma_2^2}\Big)^\tau,$$
$$U^*=\sum_{j=1}^n Y_j^2,\quad V^*=\Big(\sum_{i=1}^m X_i^2+\sum_{j=1}^n Y_j^2,\bar X,\bar Y\Big)^\tau.$$

记

$$\begin{aligned}W&:=\frac{1}{n-1}\sum_{j=1}^n(Y_j-\bar Y)^2\Big/\frac{1}{m-1}\sum_{i=1}^m(X_i-\bar X)^2\\&=\frac{(U^*-nV_3^{*2})/(n-1)}{(V_1^*-mV_2^{*2}-U^*)/(m-1)}.\end{aligned}\tag{3.26}$$

考虑第一个假设. 在给定 V^* 时 W 对 U^* 严格增, 而在 $\sigma_1=\sigma_2:=\sigma$ 时, 由 Basu 定理知 W 与 V^* 独立, 且 $1/W\sim F_{m-1,n-1}$, 故由定理 3.10 或定理 3.11, 假设 $\sigma_1^2\leqslant\sigma_2^2\leftrightarrow\sigma_1^2>\sigma_2^2$ 的水平为 α 的 UMP 无偏检验有接受域 $1/W\leqslant F_{m-1,n-1}(\alpha)$.

现考虑第二个假设. 当 $\sigma_1=\sigma_2:=\sigma$ 时, $W\sim F_{n-1,m-1}$ 且与 V^* 独立, 故由定理 3.10, 水平为 α 的 UMP 无偏检验有接受域有 $c_1\leqslant W\leqslant c_2$ 的形式. 为确定 c_1,c_2, 记 $r=\sigma_1^2/\sigma_2^2$, 且以 $F_{n-1,m-1}(\cdot)$ 和 $f_{n-1,m-1}(\cdot)$ 记 $F_{n-1,m-1}$ 变量 F 的分布函数和密度函数. 考虑到该检验的功效函数只与比值 $r=\sigma_1^2/\sigma_2^2$ 有关, 故简记为 $\beta(r)$. 有

$$1-\beta(r)=P_{\sigma_1,\sigma_2}\{c_1r\leqslant W\sigma_1^2/\sigma_2^2\leqslant c_2r\}=P(c_1r\leqslant F\leqslant c_2r).$$

由水平为 α 和无偏性, 得到

$$F_{n-1,m-1}(c_2)-F_{n-1,m-1}(c_1)=1-\alpha;$$

$$\frac{d}{dr}\int_{c_1r}^{c_2r}f_{n-1,m-1}(t)dt\Big|_{r=1}=0,\ \text{即}\ \ c_1f_{n-1,m-1}(c_1)=c_2f_{n-1,m-1}(c_2).$$

令 $W_1=\sum_{j=1}^n(Y_j-\bar Y)^2$, $W_2=\sum_{i=1}^m(X_i-\bar X)^2$. 对第二个假设, 我们也可对 $W_1/[W_1+W_2]:=h(U^*,V^*)=(U^*-nV_3^{*2})/(V_1^*-mV_2^{*2}-nV_3^{*2})$ 应用定理 3.11, 而得到相同的结果.

在应用中, 一般使用更简单的取法: $c_1=F_{n-1,m-1}(1-\alpha/2)=1/F_{m-1,n-1}(\alpha/2)$, $c_2=F_{n-1,m-1}(\alpha/2)$.

3.5 不变检验

3.5.1 问题的提法

设 $(\mathcal{X},\mathcal{B})$ 为样本空间, $\mathcal{P}=\{P_\theta,\theta\in\Theta\}$ 为样本 X 的分布族, 满足可识别性条件 (2.13): 若 $\theta_1\neq\theta_2$, 则 $P_{\theta_1}\neq P_{\theta_2}$.

类似 2.2.1 节, 引进 $\mathcal{X}$ 上的一个变换群 G, 其中每个元 g 都是 $(\mathcal{X},\mathcal{B})$ 到自身上的一个一对一的可测变换, 则由 $g(\mathcal{B})=\mathcal{B}$, 样本空间在变换群 G 之下不变.

对 $\forall\ g\in G$, 存在参数空间到自身上的一个一一变换 $\bar g$, 使对任一 $\theta\in\Theta$, 当 $X\sim P_\theta$ 时, $gX\sim P_{\bar g\theta}$, 且 $\bar G:=\{\bar g:g\in G\}$ 构成一个群, 映照 $g\mapsto\bar g$ 为 G 上的一个同态. 此时样本分布族在变换群 $\bar G$ 之下不变.

考虑形如 (3.1) 的检验问题 $H:\theta\in\Theta_0\leftrightarrow K:\theta\in\Theta_1$. 如果

$$\bar g\Theta_0:=\{\bar g\theta:\theta\in\Theta_0\}=\Theta_0,\ \ \bar g\Theta_1=\Theta_1,\ \ \text{对任一}\ \bar g\in\bar G,\tag{3.27}$$

则称**上述检验问题在变换群 G 之下不变**.

此时, 如果一个检验 ϕ 满足条件

$$\phi(gx)=\phi(x),\ \ \text{对任何}\ g\in G,x\in\mathcal{X},\tag{3.28}$$

则称 ϕ 是检验问题 (3.1) 的一个**不变检验**.

正如 2.2.1 节所述, 如果把 g 视为样本空间上的一个坐标变换, (3.28) 不过是说, 若检验问题在坐标变换下不变, 则采用的检验也应在此坐标变换下不变.

极大不变量 设 $\mathcal{G}$ 为空间 Ω 上的一个一一变换群. 在变换群 $\mathcal{G}$ 之下, 空间 Ω 被分成一些互不相交的轨道, 包含点 ω 的轨道是 $\{g\omega:g\in\mathcal{G}\}$. 在 2.2.1 节, 变换群

$\mathcal{G}$ 之下的一个不变量定义为 Ω 上的一个函数, 它在群 $\mathcal{G}$ 的每条轨道上取常数值. 如果某一不变量 $h(\omega)$ 在不同轨道上总是取不同的值, 则称其为一个极大不变量. 易见, 若 $m(\omega)$ 为极大不变量, 则 $h(\omega)$ 为一不变量当且仅当 $h(\omega)$ 可表示为 $m(\omega)$ 的函数.

依照这一术语, 检验问题 (3.1) 在群 G 下不变, 无非是说 Θ_0 和 Θ_1 包含了导出群 $\bar{G}$ 的完整轨道, 而 ϕ 为不变检验, 等同于 ϕ 为群 G 下的一个不变量.

定理 3.12 *不变检验 ϕ 的功效函数 β_ϕ 是群 $\bar{G}$ 下的不变量, 因而为群 $\bar{G}$ 下极大不变量 $v(\theta)$ 的函数.*

证明 任取 $g \in G$. 由 $\phi(X) = \phi(gX)$ 及 (2.19) 立得

$$\mathrm{E}_\theta[\phi(X)] = \mathrm{E}_\theta[\phi(gX)] = \mathrm{E}_{\bar{g}\theta}[\phi(X)].$$

这就是所要证明的. ■

由这一定理, 引进不变性缩小了参数空间, 此因不变检验的功效函数在同一轨道上取常数值, 因而在一个轨道上只需取一代表点即可. 然而更有意义的是, 不变性给入围的检验加上了限制, 此时只需考虑形如 $\phi(M(x))$ 的检验函数, 其中 $M(x)$ 为群 G 下的极大不变量, 从而给寻求最优不变检验提供了实质性的简化.

3.5.2 一致最优不变检验

设检验问题 (3.1) 在群 G 下不变. 以 $\mathcal{F}$ 记水平为 α 的所有不变检验的类, 若 $\phi \in \mathcal{F}$, 且对任一 $\psi \in \mathcal{F}$ 有

$$\beta_\psi(\theta) \leqslant \beta_\phi(\theta), \quad \text{对所有} \quad \theta \in \Theta_1,$$

则称 ϕ 为检验问题 (3.1) 的一个水平为 α 的**一致最优不变检验** (Uniformly Most Powerful Invariant Test), 简称 **UMP 不变检验**或 **UMPI 检验**.

如上所述, 不变检验就是仅依赖于极大不变量的检验. 但等价的极大不变量可以有多种不同的形式, 为寻求 UMP 不变检验, 可从中挑出一个合适的, 记为 M. 如能求得 M 的分布族并据此导出 (3.1) 的基于 M 的水平为 α 的 UMP 检验, 则它就是在 X 的分布族下 (3.1) 的水平为 α 的 UMP 不变检验. 今举例如下.

例 3.5 设 $X_1, \cdots, X_n$ 为从均匀分布总体 $\{U(-\theta, \theta), \theta > 0\}$ 抽取的 i.i.d. 样本, 任取 $\theta_0 > 0$, 考虑检验问题 $H : \theta = \theta_0 \leftrightarrow K : \theta \neq \theta_0$.

取 $(\mathbf{R}^n, \mathcal{B}_n)$ 为样本空间. 令 $g_s(x_1, \cdots, x_n) = (s_1 x_1, \cdots, s_n x_n)$, 每个 s_i 均可取 1 或 -1, $s = (s_1, \cdots, s_n)$, 所有这些 g_s 构成样本空间上的一个乘法变换群 G, 乘法定义为 $g_s g_{s^*} = g_{ss^*}$, 其中 $ss^* = (s_1 s_1^*, \cdots, s_n s_n^*)$. 易见, 对任一 $g \in G$, $\bar{g}$ 均为参数空间 $\Theta = (0, \infty)$ 上的恒等变换, 故 $\bar{G}$ 仅由一个恒等元组成. 且检验问题在群 G 和导出群 $\bar{G}$ 下不变.

极大不变量 M 可取为 $(|X_1|, \cdots, |X_n|)$, 各分量 i.i.d., $|X_1| \sim U(0, \theta), \theta > 0$. 检验问题 $H : \theta = \theta_0 \leftrightarrow K : \theta \neq \theta_0$ 的基于 M 的水平为 α 的 UMP 检验为

$$\psi(M) = \begin{cases} 1, \text{ 若 } T > \theta_0, \text{或 } T < \theta_0 \alpha^{1/n}, \\ 0, \text{ 否则.} \end{cases}$$

此处, $T=\max_{1\leqslant i\leqslant n}|X_i|$. 参看习题 3.10. ψ 即为该检验问题的基于样本 $X_1,\cdots,X_n$ 的水平为 α 的 UMP 不变检验.

但在一般情况下, 群 G 的极大不变量的分布往往难以求得, 因而上述方法难以施行, 此时一种可能的途径是借助于某个充分统计量 T. 对每个 $g\in G$, 能否在 T 的值域空间 $\mathcal{T}$ 上导出一个变换 g^*, 使得对每个 $t=T(x)$, 有 $g^*(t)=g^*[T(x)]=T(gx)$, 则是这种方法是否可行的关键.

对于任给的 g, 为要 g^* 有确切的定义, 其充要条件为

$$T(x_1)=T(x_2)\text{ 蕴含 }T(gx_1)=T(gx_2),\quad \text{对所有的 } g\in G. \tag{3.29}$$

如果 (3.29) 成立, 则容易证明, 映照 $g\mapsto g*$ 是 G 到 $G^*:=\{g^*:g\in G\}$ 上的一个同态, 作为群 G 的一个同态像, G^* 是一个群, 称为群 G 导出的 $\mathcal{T}$ 空间上的一个变换群.

在许多情形, 不难求得群 G^* 的一个极大不变量 $M=M(T)$ 和它的分布族, 如果检验 $\psi(M)$ 是基于 M 的水平为 α 的 UMP 检验, 则在一定的条件下, 即为原先的检验问题的水平为 α 的 UMP 不变检验. 此处存在的一个理论问题是, 基于充分统计量的决策函数类虽可取代全体决策函数类, 但基于充分统计量的不变决策类 (估计中用同变一词) 是否能取代全体不变决策类, 并不是理所当然的. 在讨论同变估计时也遇到过类似的问题. 此处, 我们综合了 Lehmann (1986) 的 6.5 节中的有关定理, 写成如下便于应用的定理形式. 读者如有兴趣, 可参看该书的有关论证.

定理 3.13 设 $(\mathcal{X},\mathcal{B})$ 为样本空间, $X\sim P_\theta\ll\mu,\theta\in\Theta$, μ 为 $\mathcal{B}$ 上的一个 σ-有限测度, T 是充分统计量. 设 G 为 $(\mathcal{X},\mathcal{B})$ 上的一个变换群, 它使得检验问题 (3.1) 不变, 且 (3.29) 成立. 假定 $(G,\mathcal{C})$ 为一可测空间, 且

(i) 映照 $(x,g)\mapsto gx$ 是 $(\mathcal{X}\times G,\mathcal{B}\times\mathcal{C})$ 到 $(\mathcal{X},\mathcal{B})$ 的可测映照;

(ii) 在 $(G,\mathcal{C})$ 上存在一个 σ-有限测度 ν, 使得对于 $C\in\mathcal{C}$, 有

$$\nu(C)=0\Longrightarrow\nu(Cg)=0,\quad \text{对任一 } g\in G.$$

此处 $Cg:=\{hg:h\in C\}$.

在上述条件下, 如果存在基于 T 的水平为 α 的 UMP G^* 不变检验, 则它即为检验问题 (3.1) 的基于样本 X 的水平为 α 的 UMP G 不变检验.

特别地, 设 $M=M(T)$ 是群 G^* 之下的一个极大不变量, 若存在基于 M 的水平为 α 的 UMP 检验 $\psi(M)$, 则它即为检验问题 (3.1) 的基于样本 X 的水平为 α 的 UMP 不变检验.

读者也可参看 Shao (2003) 命题 6.3.

此处我们附带指出, 如果假定在 $(G,\mathcal{C})$ 上存在一个 σ-有限测度 ν, 使得

$$\nu(C)=\nu(Cg),\quad \text{对所有 } g\in G,C\in\mathcal{C},$$

即存在一个右不变测度, 则条件 (ii) 自然满足.

例 3.6 设 $X_1,\cdots,X_n$ 为从正态总体 $\{N(\theta,\sigma^2),\theta\in\mathbf{R},\sigma>0\}$ 抽取的 i.i.d. 样本. 考虑乘法群 $G=\{g_c,\ c>0\}$, 其中 $g_c(x^n)=cx^n,x^n:=(x_1,\cdots,x_n)^\tau$. 容

易验证, 检验问题 $\theta \leqslant 0 \leftrightarrow \theta > 0$ 在群 G 下不变, g_c 在参数空间的导出变换为 $\bar{g}_c(\theta,\sigma) = (c\theta, c\sigma)$. 现求上述检验问题在群 G 下的 UMP 不变检验.

例 3.5 中的方法对本例不适用, 因极大不变量的分布难以求得. 但本例有充分统计量 $T = (\bar{X}, S)$, 其中 $S = [\sum_{i=1}^{n}(X_i - \bar{X})^2/(n-1)]^{1/2}$. 变换群 G 在 T 的值域空间 $\mathcal{T}$ 上导出一个乘法变换群 $G^* = \{g_c^* : c > 0\}$, 其中 $g_c^*(\bar{x}, s) = (c\bar{x}, cs)$. 容易验证, 定理 3.13 中的条件成立. 变换群 G^* 下的一个极大不变量为 $M = \sqrt{n}\bar{X}/S$, $M \sim t_{n-1,\delta}, \delta \in \mathbf{R}$, 其中 $\delta = \sqrt{n}\theta/\sigma$. 这样, 通过充分统计量并加上不变限制之后, 问题转化为, 要依据来自分布族 $\{t_{n-1,\delta}, \delta \in \mathbf{R}\}$ 的样本 M, 检验假设 $\delta \leqslant 0 \leftrightarrow \delta > 0$.

任取 $\delta_1 > 0$, 先考虑检验问题 $\delta = 0 \leftrightarrow \delta = \delta_1$. 为此, 以 $f(\cdot|n-1,\delta)$ 记 $t_{n-1,\delta}$ 的密度, 则由 (1.36),

$$\frac{f(m|n-1,\delta_1)}{f(m|n-1,0)} = \sum_{i=0}^{\infty} c_{ni}\delta_1^i \exp(-\delta_1^2/2)\Big(\frac{m}{\sqrt{n-1+m^2}}\Big)^i, \quad c_{ni} > 0.$$

此式为 m 的增函数, 而当 $\delta = 0$ 时 $M \sim t_{n-1}$, 故 $\delta = 0 \leftrightarrow \delta = \delta_1$ 的水平为 α 的 MP 检验 $\psi(M)$ 有否定域 $M > t_{n-1}(\alpha)$. 容易验证, 此检验的功效函数为 δ 的严增函数, 且检验的形式与 $\delta_1 > 0$ 无关, 故此检验也是检验问题 $\delta \leqslant 0 \leftrightarrow \delta > 0$ 水平为 α 的 UMP 检验. 由定理 3.13, 检验 $\psi(M)$ 也是 $\theta \leqslant 0 \leftrightarrow \theta > 0$ 在群 G 下水平为 α 的 UMP 不变检验.

例 3.7 某产品的一项质量指标 X 服从正态分布 $N(a, \sigma^2)$, 只有当 $X > u$ 时才合格, 其中 u 为已知. 不合格率为 $p = P_{a,\sigma}(X \leqslant u)$. 现有 X 的 i.i.d. 样本 $X_1, \cdots, X_n$, 要检验假设 $H : p \geqslant p_0 \leftrightarrow p < p_0$. 通过一个平移变换, 就可把问题化为 $u = 0$ 的情形. 故不妨取 $u = 0$. 这时 $p = \Phi(-a/\sigma)$, 而 $p \geqslant p_0$ 转化为 $\theta := a/\sigma \leqslant \Phi^{-1}(1-p_0) := \theta_0$, 其中 Φ 和 Φ^{-1} 分别为标准正态的分布函数及其反函数. 在例 3.6 的乘法群 G 之下, 检验问题 $\theta \leqslant \theta_0 \leftrightarrow \theta > \theta_0$ 是不变的. 转到充分统计量 $T = (\bar{X}, S)$, 得极大不变量 $M = \sqrt{n}\bar{X}/S$, 其分布为 $\{t_{n-1,\delta} : \delta \in \mathbf{R}\}$, 其中 $\delta = \sqrt{n}\theta$. 检验问题转化为: 基于样本 M 检验 $\delta \leqslant \delta_0 \leftrightarrow \delta > \delta_0$. 可以证明, $\{t_{n-1,\delta} : \delta \in \mathbf{R}^1\}$ 为关于 M 的单调似然比族, 故水平为 α 的 UMP 检验存在, 其否定域为 $M > t_{n-1,\delta_0}(\alpha)$, 其中 $\delta_0 = \sqrt{n}\theta_0$. 由定理 3.13, 此检验即为原检验问题在群 G 下水平为 α 的 UMP 不变检验.

关于 $\{t_{n-1,\delta} : \delta \in \mathbf{R}\}$ 为单调似然比族的证明有一定难度, 读者如有兴趣, 可参看 Lehmann (1986), 第 295—296 页.

3.6 习题和补充

在以下陈述中, $I(A)$ 表示集合 A 的示性函数, $0 < \alpha < 1$ 表示指定的检验水平,

$$X_{(1)} := \min_{1\leqslant i\leqslant n} X_i, \quad X_{(n)} := \max_{1\leqslant i\leqslant n} X_i,$$

对于小写字母也有类似的记号.

3.1 设参数空间为一个一维开区间, $\{dP_\theta(x) = f_\theta(x)d\mu\}$ 关于 $T(x)$ 为一 MLR 族, 则对 T 的任一严格增 (非降) 函数 $g(T)$, $\mathrm{E}_\theta g[T(X)]$ 为 θ 的严格增 (非降) 函数.

3.2 设样本空间 $(\mathcal{X}, \mathcal{B})$ 是欧氏的, 且 P_0, P_1 关于 Lebesgue 测度有密度, 则对任给的水平 α, 检验问题 "$P_0 \leftrightarrow P_1$" 存在非随机化的 MP 检验.

3.3 设 $X_1, \cdots, X_n$ 为从均值已知的正态总体 $\{N(a, \sigma^2), \sigma > 0\}$ 中抽取的 i.i.d. 样本. 给定 $\sigma_0 > 0$. 试求检验问题 $H : \sigma \leqslant \sigma_0 \leftrightarrow \sigma > \sigma_0$ 的水平为 α 的 UMP 检验.

3.4 设 $X_1, \cdots, X_n$ 为从分布族 $\{E(0, \theta) : \theta > 0\}$ 中抽取的 i.i.d. 样本. 给定 $\theta_0 > 0$. 试求检验问题 $H : \theta \leqslant \theta_0 \leftrightarrow \theta > \theta_0$ 的水平为 α 的 UMP 检验.

3.5 设样本 $X_1, \cdots, X_n$ i.i.d., 其共同分布为

$$P_\theta(X_1 = 0) = \theta^2, \quad P_\theta(X_1 = 1) = 2\theta(1-\theta), \quad P_\theta(X_1 = 2) = (1-\theta)^2,$$

其中, $0 < \theta < 1$. 以 N_j 记样本中等于 j 的个数, $j = 0, 1, 2$. 令 $L(\theta)$ 为 θ 的似然函数.

(1) 证明 $\{L(\theta), 0 < \theta < 1\}$ 是关于 $(2N_0 + N_1)$ 的单调似然比分布族.

(2) 对给定的 $0 < \lambda < 1$, 求检验问题 "$H : \mathrm{E}_\theta(X_1) \leqslant 2\lambda \leftrightarrow K : \mathrm{E}_\theta(X_1) > 2\lambda$" 的水平为 α 的 UMP 检验 (不必确切求有关的常数).

(3) 求 $(2N_0 + N_1)$ 的分布.

3.6 设 $(X, Y) \sim N(0, 0; \sigma_1^2, \sigma_2^2, \rho)$, 而 $(X_1, Y_1), \cdots, (X_n, Y_n)$ 是取自这一总体的一个随机样本.

(1) 证明 $U := X + Y$ 和 $V := X - Y$ 独立当且仅当 $\sigma_1^2 = \sigma_2^2$. 因此, 为要检验假设 $H : \sigma_1 = \sigma_2$, 我们可以检验等价的假设 $H^* : \rho^* = 0$, 此处 ρ^* 为 (U, V) 的相关系数. 令 $U_i = X_i + Y_i, V_i = X_i - Y_i, i = 1, \cdots, n$, 且用 r^* 记 (U, V) 的样本相关系数.

(2) 基于 r^* 定义一个检验统计量, 服从自由度为 $n-2$ 的学生氏 t-分布, 且为 r^* 的单调函数.

3.7 设 $X_1, \cdots, X_n$ i.i.d., $X_1 \sim E(0, \theta), \theta > 0$. 在这些数据中, 假定我们只能观测到前 r 个次序样本 $X_{(1)}, X_{(2)}, \cdots, X_{(r)}$, 其中 $1 < r < n$. 在这一定数删失模型下试求检验问题 $H : \theta \leqslant \theta_0 \leftrightarrow K : \theta > \theta_0$ 的水平为 α 的 UMP 检验. 此处, θ_0 为一给定的正数.

3.8 设 X 为从密度族 $\{f(x|\theta), -\infty < \theta < \infty\}$ 中抽取的样本,

$$f(x|\theta) = \begin{cases} (1/v)I(0 \leqslant x \leqslant v), & \text{若 } \theta = 0, \\ \theta \mathrm{e}^{-\theta x}/(1 - \mathrm{e}^{-\theta v})I(0 \leqslant x \leqslant v), & \text{若 } \theta \neq 0 \end{cases}$$

(注意 θ 可取负值). 给定 $\theta_0 > 0$. 试给出假设检验问题 "$H : \theta \leqslant \theta_0 \leftrightarrow K : \theta > \theta_0$" 的水平为 α 的 UMP 检验, 并证明你的结论.

3.9 设 X 有密度 $\{f_\theta(x) = (1 + x - \theta)^{-2} I(x \geqslant \theta), \theta_0$ 为一给定的实数. 证明: 否定域为 $\{X < \theta_0\} \cup \{X > \theta_0 - 1 + 1/\alpha\}$ 的检验为检验问题 "$H_0 : \theta = \theta_0 \leftrightarrow H_1 : \theta < \theta_0$"

的水平为 α 的 UMP 检验.

3.10 设 $X_1,\cdots,X_n$ 为从均匀分布族 $\{U(0,\theta),\theta>0\}$ 中抽取的 i.i.d. 样本. 给定 $\theta_0>0$. 则对检验问题 "$H:\theta=\theta_0 \leftrightarrow K:\theta\neq\theta_0$", 存在水平为 α 的一个 UMP 检验, 它可表述为:

$$\phi(x_{(n)})=\begin{cases}1, & \text{若 } x_{(n)}>\theta_0 \text{ 或 } x_{(n)}<\theta_0\alpha^{1/n},\\ 0, & \text{否则.}\end{cases}$$

[双边假设存在 UMP 检验的例子极为罕见, 本题和下面两题给出了这样的例子.]

3.11 设 $X_1,\cdots,X_n$ 为从指数分布族 $\{E(a,1),a\in\mathbf{R}\}$ 中抽取的 i.i.d. 样本. 任给一个实数 a_0. 求检验问题 "$H:a=a_0\leftrightarrow a\neq a_0$" 的水平为 α 的 UMP 检验.

3.12 设 $X_1,\cdots,X_n$ 为取自总体 X 的 i.i.d. 样本, X 服从 $\{1,\cdots,\theta\}$ 上的均匀分布, 其中 θ 为一未知的自然数. 假定 θ_0 为一已知的自然数. 证明:

(1) 对于检验问题 "$H:\theta\leqslant\theta_0\leftrightarrow K:\theta>\theta_0$", 检验

$$\phi_1(x_{(n)})=\begin{cases}1, & \text{若 } x_{(n)}>\theta_0,\\ \alpha, & \text{若 } x_{(n)}\leqslant\theta_0\end{cases}$$

是一个水平为 α 的 UMP 检验.

(2) 对于检验问题 "$H:\theta=\theta_0\leftrightarrow K:\theta\neq\theta_0$", 检验

$$\phi(x_{(n)})=\begin{cases}1, & \text{若 } x_{(n)}>\theta_0 \text{ 或 } x_{(n)}\leqslant\theta_0\alpha^{1/n},\\ 0, & \text{否则}\end{cases}$$

是一个水平为 α 的 UMP 检验.

[提示: 证明在 $\theta>\theta_0$ 处, 检验 ϕ 与第 1 问中的 UMP 检验 ϕ_1 有同样的功效; 设 ϕ_2 是检验问题 "$\theta=\theta_0\leftrightarrow\theta<\theta_0$" 水平为 α 的 UMP 检验, 再证明在 $\theta<\theta_0$ 处, 检验 ϕ 与 ϕ_2 有同样的功效即可.]

3.13 设 $X_1,\cdots,X_n$ 为从均匀分布族 $\{U(\theta,2\theta),\theta>0\}$ 中抽取的 i.i.d. 样本. 给定 $\theta_0>0$. 问检验问题 "$H:\theta\leqslant\theta_0\leftrightarrow K:\theta>\theta_0$" 水平为 α 的 UMP 检验是否存在?

[提示: 取 $\theta_1>\theta_0$. 考虑检验问题 "$\theta=\theta_0\leftrightarrow\theta=\theta_1$", 证明得到的水平为 α 的 MP 检验也是 "$\theta\leqslant\theta_0\leftrightarrow\theta=\theta_1$" 的水平为 α 的 UMP 检验. 再考察此检验是否与 θ_1 有关.]

3.14 设 $X_1,\cdots,X_n$ 为从均匀分布族 $\{U(\theta,\theta+1),-\infty<\theta<\infty\}$ 中抽取的 i.i.d. 样本, $n\geqslant 2$. 证明对于检验问题 $H:\theta\leqslant 0\leftrightarrow K:\theta>0$, 存在一个水平为 α 的 UMP 检验, 当 $X_{(1)}>C(\alpha)$ 或 $X_{(n)}>1$ 时否定原假设. 确定待定常数 $C(\alpha)$.

*3.15 给定 $\theta_1<\theta_2$. 在定理 3.4 的单参数指数族中检验假设

$$H:\theta\leqslant\theta_1 \text{ 或 } \theta\geqslant\theta_2\leftrightarrow K:\theta_1<\theta<\theta_2.$$

证明: 若函数 $Q(\cdot)$ 严格增, 则存在一个水平为 α 的 UMP 检验 ϕ, 它可写成下

面的形式:

$$\phi(x)=\begin{cases}1, & \text{若 } c_1<T(x)<c_2\ (c_1<c_2),\\ \gamma_i, & \text{若 } T(x)=c_i,\ i=1,2,\\ 0, & \text{若 } T(x)<c_1 \text{ 或 } T(x)>c_2,\end{cases}$$

其中 $c_1,c_2,\gamma_1,\gamma_2$ 由 $\mathrm{E}_{\theta_1}[\phi(X)]=\mathrm{E}_{\theta_2}[\phi(X)]=\alpha$ 决定.

3.16 **用 Bayes 方法求 UMP 检验**.

设 $(\mathcal{X},\mathcal{B})$ 为样本空间, $X\sim f(x,\theta)d\mu(x),\theta\in\Theta$. 为了求得检验问题

$$H:\theta\in\Theta_H\leftrightarrow K:\theta\in\Theta_K\equiv\Theta-\Theta_H$$

的水平为 α 的 UMP 检验, 一般的步骤是任取一点 $\theta_1\in\Theta_K$, 考虑检验问题

$$H:\theta\in\Theta_H\leftrightarrow K_1:\theta=\theta_1.$$

如果求得的该问题的水平为 α 的 UMP 检验与 $\theta_1\in\Theta_K$ 无关, 则它显然就是检验问题 "$H\leftrightarrow K$" 的水平为 α 的 UMP 检验.

在某些情况下, 可以应用 Bayes 方法求得检验问题 "$H\leftrightarrow K_1$" 的 UMP 检验. 为此, 我们需要在 Θ 中引进 σ-域 $\mathcal{B}_\Theta$, 并假定 $f(x,\theta)$ 为 $\mathcal{B}\times\mathcal{B}_\Theta$-可测. 设 λ 为适当选取的 Θ_H 上的一个先验概率测度. 令

$$h_\lambda(x)=\int_{\Theta_H}f(x,\theta)d\lambda(\theta),$$

不难看出, h_λ 为 $\mathcal{B}$-可测, 且 $h_\lambda(x)d\mu(x)$ 为 $(\mathcal{X},\mathcal{B})$ 上的一个概率测度. 考虑检验问题

$$H_\lambda:X\sim h_\lambda(x)d\mu(x)\leftrightarrow K_1:X\sim P_{\theta_1}.$$

则不难证明下述定理:

以 ϕ_λ 记检验问题 "$H_\lambda\leftrightarrow K_1$" 的水平为 α 的 UMP 检验. 若 ϕ_λ 对 "$H\leftrightarrow K_1$" 也是水平为 α 的检验, 则它是 "$H\leftrightarrow K_1$" 的水平为 α 的 UMP 检验, 且若 ϕ_λ 是 "$H_\lambda\leftrightarrow K_1$" 的唯一的水平为 α 的 UMP 检验, 则它也是 "$H\leftrightarrow K_1$" 的唯一的水平为 α 的 UMP 检验.

读者可以把这一定理的证明作为一个练习. [参看陈希孺 (1981, 1997) 中的定理 3.2.3、例 3.2.4, 以及第三章习题 5 和 6.]

3.17 设 $X\sim P_i,i\in\{0,1,2\}$, P_i 如下所示:

	1	2	3	4
P_0	2/13	4/13	3/13	4/13
P_1	4/13	2/13	1/13	6/13
P_2	4/13	3/13	2/13	4/13

求检验问题 "$H:\{P_0,P_1\}\leftrightarrow K:P_2$" 的水平 $\alpha=5/13$ 的 UMP 检验.

3.18 设 $X_1,\cdots,X_n$ 为从正态总体 $\{N(0,\theta),\theta>0\}$ 中抽取的 i.i.d. 样本. 给定 $\theta_0>0,0<\theta_1<\theta_2$, 试求检验问题 "$H_1:\theta_1\leqslant\theta\leqslant\theta_2\leftrightarrow K_1:\theta\notin[\theta_1,\theta_2]$" 和

"$H_2:\theta=\theta_0 \leftrightarrow K_2:\theta\neq\theta_0$" 的水平为 α 的 UMP 无偏检验.

3.19 设 $X_1,\cdots,X_n$ 为从 Poisson 分布总体 $\{P(\theta),\theta>0\}$ 中抽取的 i.i.d. 样本. 给定 $\theta_0>0$ 和 $0<\theta_1<\theta_2$, 试求检验问题 $H_1:\ \theta_1\leqslant\theta\leqslant\theta_2 \leftrightarrow K_1:\ \theta\notin[\theta_1,\theta_2]$ 和检验问题 $H_2:\ \theta=\theta_0\leftrightarrow K_2:\ \theta\neq\theta_0$ 的水平为 α 的 UMP 无偏检验.

3.20 设 $X_1,\cdots,X_n$ 是来自分布 $f(x,\theta)dx=\theta x^{\theta-1}I(0<x<1)dx$ 的 i.i.d. 样本, 其中 $\theta>0$ 未知. 试求检验问题 $H_1:\theta_1\leqslant\theta\leqslant\theta_2\leftrightarrow K_1:\theta\notin[\theta_1,\theta_2]$ 和检验问题 $H_2:\theta=\theta_0\leftrightarrow K_2:\ \theta\neq\theta_0$ 的水平为 α 的 UMP 无偏检验.

3.21 设样本 X 有多参数指数族分布 (3.22):

$$dP^X_{\theta,\vartheta}(x)=C(\theta,\vartheta)\exp\{\theta U(x)+\vartheta^\tau V(x)\}d\mu(x),\quad(\theta,\vartheta^\tau)^\tau\in\Theta,$$

此处 Θ 是自然参数空间, Θ 在 $\mathbf{R}^{d+1}$ 中的内点集 Θ^{o} 非空. 给定 $\theta_1<\theta_2$, 假定 $\mathbf{R}^{d+1}$ 中的超平面 $\theta=\theta_i$ 与 Θ^{o} 有非空的交, $i=1,2$. 证明对检验问题 $H:\theta\leqslant\theta_1$或 $\theta\geqslant\theta_2\leftrightarrow K:\theta_1<\theta<\theta_2$, 存在一个水平为 α 的 UMP 无偏检验 $\phi(U,V)$, 它具有下面的形式:

$$\phi(u,v)=\begin{cases}1, & \text{若 } c_1(v)<u<c_2(v)\ (c_1(v)<c_2(v)),\\ \gamma_i(v), & \text{若 } u=c_i(v),\ i=1,2,\\ 0, & \text{若 } u<c_1(v) \text{ 或 } u>c_2(v),\end{cases}$$

其中 $c_i(v),\gamma_i(v),i=1,2$ 由 $\mathrm{E}_{\theta_1}[\phi(U,V)|v]=\mathrm{E}_{\theta_2}[\phi(U,V)|v]=\alpha$ 决定.

*3.22 设 $X_1,\cdots,X_m$ 和 $Y_1,\cdots,Y_n$ 分别为来自 $N(a,\sigma^2)$ 和 $N(a,\rho\sigma^2)$ 的 i.i.d. 样本, 两样本整体独立, $\rho>0$ 已知, a 和 σ 未知. 试求检验问题 $H:a=a_0\leftrightarrow K:a\neq a_0$ 的水平为 α 的 UMP 无偏检验, 此处 a_0 为一指定的实数.

3.23 设 $X_1,\cdots,X_n$ 是从分布为 $E(a,b):-\infty<a<\infty,b>0$ 的总体中取出的 i.i.d. 样本. 记 $T=nX_{(1)},V=2\sum_{i=2}^n(X_{(i)}-X_{(1)})$, 则两者独立, 且 (T,V) 为充分统计量. 分别写出它们的分布, 并求检验问题 $H:b=1\leftrightarrow K:b\neq1$ 的水平为 α 的 UMP 无偏检验.

[参看 1.6.2 节例 1.13.]

3.24 设 X 和 Y 是来自 Poisson 分布 $P(\lambda)$ 和 $P(\mu)$ 的独立样本,

$$P(X=x,\ Y=y)=\frac{\mathrm{e}^{-(\lambda+\mu)}}{x!y!}\exp\Big[y\log\frac{\mu}{\lambda}+(x+y)\log\lambda\Big],$$

$\lambda>0,\mu>0$ 都未知. 试应用定理 3.10, 求检验问题 $H_1^*:\mu\leqslant\lambda\leftrightarrow K_1^*:\mu>\lambda$ 和 $H_4^*:\mu=\lambda\leftrightarrow K_4^*:\mu\neq\lambda$ 的水平为 α 的 UMP 无偏检验.

3.25 设样本 X 和 Y 独立, $X\sim B(m,p_1),Y\sim B(n,p_2)$, m,n 已知, $p_i=1-q_i$ 均取值于 (0, 1) 区间, p_1,p_2 未知.

$$\begin{aligned}P(X=x,\ Y=y)&=\binom{m}{x}\binom{n}{y}p_1^xq_1^{m-x}p_2^yq_2^{n-y}\\&=\binom{m}{x}\binom{n}{y}q_1^mq_2^n\exp\Big[y\Big(\log\frac{p_2}{q_2}-\log\frac{p_1}{q_1}\Big)+(x+y)\log\frac{p_1}{q_1}\Big].\end{aligned}$$

与定理 3.10 类似, 可以考虑关于参数 $\theta\equiv\log\rho$ 的四个假设检验问题: $H_i\leftrightarrow K_i$,

$i=1,2,3,4$. 其中

$$\rho:=\frac{p_2}{q_2}\Big/\frac{p_1}{q_1}$$

称为交比 (odds ratio). 试求其中两个检验问题 $H_1^*:p_2\leqslant p_1\leftrightarrow K_1^*:p_2>p_1$ 和 $H_4^*:p_2=p_1\leftrightarrow K_4^*:p_2\neq p_1$ 的水平为 α 的 UMP 无偏检验.

3.26 设 $X_1,\cdots,X_n$ 独立, X_k 的分布为

$$P(X_k=1)=1-P(X_k=0)=\frac{1}{\exp(-\theta_1-k\theta_2)},$$

$\theta_1\geqslant 0,\theta_2\geqslant 0$ 为未知参数, 求检验问题 "$\theta_2=0\leftrightarrow\theta_2>0$" 的水平为 α 的 UMP 无偏检验.

3.27 设 $X_1,\cdots,X_n$ 独立, 其中 $X_k\sim N(k\mu,\sigma^2),k=1,\cdots,m;\ \ X_j\sim N(0,\sigma^2),\ j=m+1,\cdots,n$, 而 μ 和 σ^2 为未知参数. 求检验问题 "$\mu=0\leftrightarrow\mu\neq 0$" 的水平为 α 的 UMP 无偏检验.

3.28 设 X_1,X_2 独立且均为 0, 1 两值变量, 取值 1 的概率分别为 p_1 和 p_2, 参数空间为 $\Theta=\{(p_1,p_2):0<p_1,p_2<1\}$. 设 $(a,b)\in\Theta$. 证明: 假设检验问题 "$H:p_1=a,p_2=b\leftrightarrow(p_1,p_2)\neq(a,b)$" 仅有的一个水平为 α 的无偏检验为 $\phi(x_1,x_2)\equiv\alpha$.

[参看陈希孺 (1999, 2009) 中的第五章习题 35.]

*3.29 设 X 和 Y 为独立的一维指数族变量, 其联合分布为

$$(X,Y)\sim C(\theta_1)\mathrm{e}^{\theta_1T(x)}d\mu(x)K(\theta_2)\mathrm{e}^{\theta_2U(y)}d\nu(y).$$

假定 (a,b) 为自然参数空间 Θ 的一个内点, 且以概率 1, T 和 U 中的每一个变量至少取到三个值. 证明: 假设检验问题 "$H:\theta_1=a,\theta_2=b\leftrightarrow(\theta_1,\theta_2)\neq(a,b)$" 的水平为 α 的 UMP 检验不存在.

[此题和下一题提供了两个不同的例子, 取自 Lehmann (1986), 第 171 页. 请读者注意第 171—172 页每个题后仔细的提示.]

*3.30 设样本 $(X,Y)\sim C(\theta_1,\theta_2)\exp(\theta_1x+\theta_2y)d\mu(x,y),(\theta_1,\theta_2)\in\Theta$, 而 (a,b) 为 Θ 在 $\mathbf{R}^2$ 中的内点. 证明: 检验问题 "$H:\theta_1\leqslant a,\theta_2\leqslant b\leftrightarrow K:\theta_1>a$ 或 $\theta_2>b$" 仅有的一个无偏检验是 $\phi(x,y)\equiv\alpha$.

3.31 设 $X_1,\cdots,X_n$ 是从均值、方差均为未知的正态总体 $N(\mu,\sigma^2)$ 的总体中取出的 i.i.d. 样本. $G=\{g_c,c\neq 0\}$ 为 $\mathbf{R}^n$ 到自身上的一个变换群, 其中 $g_c(x_1,\cdots,x_n)=(cx_1,\cdots,cx_n)$.

(1) 证明检验问题 "$H:\mu=0\leftrightarrow K:\mu\neq 0$"在此变换群下是不变的.

(2) 求此检验问题的水平为 α 的 UMP 不变检验.

3.32 设 X_1,X_2 独立, 分别服从均值未知的正态 $N(\theta_1,1/2)$ 和 $N(\theta_2,1/2)$. 要检验假设 "$\theta_1=\theta_2\leftrightarrow\theta_1\neq\theta_2$". 考虑变换群 $G=\{g_{a,b},a=\pm1,b\in\mathbf{R}\}$, 此处, $g_{a,b}(x_1,x_2)=(ax_1+b,ax_2+b)$.

(1) 证明: 该检验问题在此变换群下是不变的, $Z=|X_1-X_2|$ 是极大不变的,

其密度仅通过 $\lambda=|\theta|$ 依赖于 θ, 写出 Z 的密度.

(2) 原问题可简化为在 Z 的分布族中检验假设 "$\lambda=0\leftrightarrow\lambda>0$", 由此求原检验问题的水平为 α 的 UMP 不变检验.

*3.33 设 $X_1,\cdots,X_n$ i.i.d, $X_1\sim U(\theta,\theta+\delta):-\infty<\theta<\infty,\delta>0$.

(1) 考虑假设检验 $H:\theta\leqslant 0\leftrightarrow K:\theta>0$. 证明在变换群 $G=\{g_c:g_cx=cx,c>0\}$ 之下, 该检验问题不变.

(2) 证明基于完全充分统计量的极大不变量为 $R=X_{(1)}/X_{(n)}$. 而对 $\bar{G}$ 而言, $\rho=\theta/\delta$ 是极大不变的.

(3) 求 $H\leftrightarrow K$ 的水平为 α 的 UMP 不变检验.

[注: $\rho\geqslant 0$ 时, R 的密度为

$f_\rho(r)=(n-1)(1-r)^{n-2}\{(1+\rho)^n-(\rho/r)^n\}I\big(\rho/(1+\rho)<r<1\big)$.]

3.34 设样本 $X\sim p_\theta(x)dx,\theta=0,\ 1$. 考虑检验问题 "$H:\theta=0\leftrightarrow K:\theta=1$". 假定事先已知 $P(\theta=0)=\pi=1-P(\theta=1)$, 且犯两种错误的损失均为 1.

(1) 证明使用某检验 ϕ 引起的 Bayes 全面风险为

$$\pi\mathrm{E}_0[\phi(X)]+(1-\pi)\mathrm{E}_1\big[1-\phi(X)\big].$$

(2) 确定一个检验使上述风险达到最小.

第四章

区间估计

4.1 基本概念

本章要在点估计的基础上讨论区间估计的有关问题. 为此, 设 $(\mathcal{X}, \mathcal{B})$ 为样本空间, $\{P_\theta, \theta \in \Theta\}$ 为分布族, 参数空间 Θ 为 $\mathbf{R}^k$ 或它的一个 Borel 子集, 样本为 X.

如前所述, 点估计就是用一个 (k-维) 估计量 $\delta(X)$ 去估计 (k-维) 未知参数 θ. 这种估计的缺点是无法从估计值看出它的精度有多大. 虽然你可以用均方误差之类的指标去刻画这种精度, 但这毕竟还是间接的. 更为直接的方法, 就是本章所要讨论的一维参数的区间估计, 以及更一般的, 向量参数的集估计. 以一维参数的估计为例, 就是要指出一个误差限 $d(X)$, 而把估计写成 $\delta(X) \pm d(X)$ 的形式. 在各种应用场合常常可以看到这种写法. 事实上这就是一种区间估计, 即估计 θ 落在区间 $[\delta(X) - d(X), \delta(X) + d(X)]$ 内. 这种表示估计精度的方法既简明又直观.

参数 θ 的一个**集估计**是指 $\mathcal{X}$ 到 Θ 的某个子集类中的一个映照 $x \mapsto S(x)$, 每当有了样本 x, 估计 θ 就落在 $S(x)$ 内. 在一维情形, 通常假定参数空间 Θ 为一个有限或无限的区间.

衡量一个集估计的优良性有两个方面的指标, 就是**可靠度**和**精度**. 首先介绍可靠度的概念.

可靠度 可靠度以 $S(X)$ 包含真参数 θ 的概率, 即 $P_\theta\{\theta \in S(X)\}$ 来衡量. 如果

$$P_\theta\{\theta \in S(X)\} \geqslant 1 - \alpha, \quad \text{对每个 } \theta \in \Theta,$$

则称 $S(X)$ 是 θ 的**置信水平**为 $1-\alpha$ 的一个**置信集**, 或者称 $1-\alpha$ 为 $S(X)$ 的一个**置信水平**. 显然置信水平不唯一: 若 $1-\alpha$ 为置信水平, $0 \leqslant 1-\gamma \leqslant 1-\alpha$, 则 $1-\gamma$ 也是.

最大的置信水平, 即 $\inf\{P_\theta\big(\theta \in S(X)\big) : \theta \in \Theta\}$, 称为 $S(X)$ 的**置信系数**.

在 $P_\theta\{\theta \in S(X)\}$ 与参数 θ 无关的情形, 我们称 S 为**相似置信集**. 这时共同的 $P_\theta\{\theta \in S(X)\}$ 即为 S 的置信系数.

从理论上讲, 要使上述概率有定义, 形如 $\{\theta \in S(x)\}$ 的集合都必须是 $\mathcal{B}$-可测的. 给出满足这种要求的普遍性条件并无必要, 因为当 $S(x)$ 具有常见的一些规则形状时, 上述可测性一般都不难验证.

实用中 $S(x)$ 通常都有简单规则的形状, 例如在 θ 为一维时, 通常考虑的 $S(x)$ 只有以下三种形式:

$$[A(x), B(x)], \quad (-\infty, C(x)] \quad \text{和} \quad [C(x), \infty),$$

其中 $A(x) \leqslant B(x), x \in \mathcal{X}$. 当我们关心某种物质所含杂质的百分比, 或关心某种材料的强度时, 就会遇到第二或第三种情形. 第一种情形称 $[A(x), B(x)]$ 为 θ 的一个**区间估计**(或**置信区间**). "置信" 一词与上面定义的置信水平有关. 在后两种情形中, $C(x)$ 分别称为 θ 的一个 **置信上限**和**置信下限**.

当 $\theta = (\theta_1, \cdots, \theta_k)^\tau$ 的维数 $k > 1$ 时, $S(x)$ 一般可取为有界或一端无界的长方体、球和椭球等, 统称为 θ 的**置信集**. 又因 $S(x)$ 一般是一个区域, 故又常称为 θ 的**区域估计**或**置信(区)域**. 形如 $\{A_i(x) \leqslant \theta_i \leqslant B_i(x), 1 \leqslant i \leqslant k\}$ 的长方形置信域又可视为对 $\theta_1, \cdots, \theta_k$ 的**同时区间估计**, 类似地有**同时置信 (上、下) 限**.

精度 粗略地说, 精度方面的要求就是希望 $S(x)$ 不要太 "大". 从不同的角度出发, 精度的界定可以有不同的方法.

一种常用的方法是引进一个合适的测度空间 $(\Theta, \mathcal{B}_\Theta, \nu)$, 要求 $S(x) \in \mathcal{B}_\Theta$, 并且 $\nu(S(x))$ 尽可能小. 例如, 通常可要求 $S(x)$ 的长度、面积或体积尽量小. 因此, 若 θ_0 为真参数, 则 $\mathrm{E}_{\theta_0}[\nu(S(X))]$ 可取为衡量置信集平均精度的一个尺度.

另一种方法是, 当参数真值为 θ_0 时要求 $S(X)$ 尽量不要包含异于 θ_0 的参数值. 对于置信限, 还要作以下补充说明: 以置信上限为例, 若 $(-\infty, S(X)]$ 包含真参数 θ_0, 它必包含所有更小的 θ 值, 故我们只需要求 $(-\infty, S(X)]$ 尽量不要包含大于 θ_0 的参数值. 由此导出的优良性准则, 使得区间估计的优良性理论, 可以很容易地通过与假设检验的联系建立起来.

显而易见, 可靠度和精度是此长彼消的两个指标: 增加置信集的可靠度, 势必会降低它的精度; 反之亦然. 为了统筹这两方面的要求, 现在普遍接受的一种做法是: 首先要求置信水平满足一定的要求, 在此限制下, 尽量寻求精度较高的置信集. 这就是 Neyman 在发展置信集优良性理论中引进的原则. 这类似于在假设检验中, 限定检验的水平即第一类错误概率的做法.

本章主要研究一维参数的区间估计的有关理论. 区间估计是一种重要的统计推断形式, 也是理论上各家争论较多的一个统计领域. 争论的焦点在于, 对区间估计问题的意义应如何理解. 这种理解的不同, 也导致了方法上的差异. 就目前情况而论, 占主导地位的, 还是 Neyman 在 1934—1938 年间引进的置信区间理论, 参看 Neyman

(1937). 此外, 我们还会对 Fisher 关于区间估计的观点, 即信仰推断法作一简介, 参看 Fisher (1930) (1935) (1973). 本章主要参考了陈希孺 (1999, 2009), 和茆诗松, 王静龙, 濮晓龙 (1998), 也引用了其中的一些例题和习题, 后面不再一一说明.

置信水平和置信系数的概念是 Neyman 区间估计理论的基本概念, 这些概念允许一种频率的解释, 例如, 当提及 $[A(X), B(X)]$ 是 θ 的一个相似置信区间时, 其置信系数为 $1-\alpha$ 是指, 如果这个估计重复使用的次数足够大, 则可以期望所得区间包含真参数 θ 的频率接近于 $100(1-\alpha)\%$.

以后我们常把置信水平是 $1-\alpha$ 的置信区间简称为 $1-\alpha$ 置信区间.

4.2 构建区间估计的方法

4.2.1 枢轴变量法

先看一个简单的例子. 设 $X_1, \cdots, X_n$ 为抽自正态总体 $N(\theta, \sigma^2)$ 的 i.i.d. 样本, $\sigma^2 > 0$ 已知, 要求构建 θ 的置信区间.

由第二章, 样本均值 $\bar{X}$ 是 θ 的一个良好的点估计, $\bar{X}$ 有分布

$$\sqrt{n}(\bar{X}-\theta)/\sigma \sim N(0,1).$$

由此可得 $P_\theta\{|\sqrt{n}(\bar{X}-\theta)| \leqslant u_{\alpha/2}\sigma\} = 1-\alpha$. 此式可改写为

$$P_\theta\{\bar{X} - u_{\alpha/2}\sigma/\sqrt{n} \leqslant \theta \leqslant \bar{X} + u_{\alpha/2}\sigma/\sqrt{n}\} = 1-\alpha. \tag{4.1}$$

此处, $u_{\alpha/2}$ 是标准正态分布的上 $\alpha/2$ 分位数. 此式指出,

$$[\hat{\theta}_1, \hat{\theta}_2] := [\bar{X} - u_{\alpha/2}\sigma/\sqrt{n}, \bar{X} + u_{\alpha/2}\sigma/\sqrt{n}]$$

是 θ 的一个 $(1-\alpha)$ 相似置信区间. 若要作 θ 的置信限, 则由 $P_\theta\{\sqrt{n}(\bar{X}-\theta) \geqslant -u_\alpha\sigma\} = 1-\alpha$ 和 $P_\theta\{\sqrt{n}(\bar{X}-\theta) \leqslant u_\alpha\sigma\} = 1-\alpha$, 知 $\bar{X} + u_\alpha\sigma/\sqrt{n}$ 和 $\bar{X} - u_\alpha\sigma/\sqrt{n}$ 分别为 θ 的 $(1-\alpha)$ 相似置信上、下限.

由此可以归纳出一种构建置信区间的一般方法: 设法找出样本 X 和参数 θ 的一个函数 $G(X,\theta)$, 其分布已知, 则对于任给的 $0 < \alpha < 1$, 可找到 $c(\theta)$ 和 $d(\theta)$, $c(\theta) < d(\theta)$, 使事件 $A_\theta := \{c(\theta) \leqslant G(X,\theta) \leqslant d(\theta)\}$ 的概率

$$P_\theta(A_\theta) \geqslant 1-\alpha, \quad \forall\ \theta \in \Theta.$$

如果对给定的 X, 集合 $J(X) = \{\theta : X \in A_\theta\}$ 总是一个有限区间, 则 $J(X)$ 即为 θ 的一个 $1-\alpha$ 置信区间. 若 $J(X)$ 总是形如 $(-\infty, C(X)]$ 的区间 (或者总是形如 $[C(X), \infty)$ 的区间), 则 $C(X)$ 即为 θ 的一个 $1-\alpha$ 置信上限 (或置信下限). **一般说来, 常可选取变量 $G(X,\theta)$, 其分布与 θ 无关, 因而 $c(\theta), d(\theta)$ 均为常数, 得到的置信区间和置信限可以是相似的, 此时 $G(X,\theta)$ 称为枢轴变量**. 枢轴变量法也由此得名.

正如上面的例子中指出的, **如果先找出 θ 的一个优良的点估计 $T = T(X)$, 通过 T 我们就可以更方便地构建枢轴变量 $G^*(T,\theta) \equiv G(X,\theta)$. 起初选用的点估计不同, 得到的枢轴变量也往往不一样.**

例 4.1 设 $X_1,\cdots,X_n$ 为抽自正态总体 $N(\mu,\sigma^2)$ 的 i.i.d. 样本, μ 和 $\sigma^2>0$ 均为未知. 给定 $\alpha\in(0,1/2)$, 要求构建 μ 和 σ^2 的 $1-\alpha$ 置信区间和置信限.

设 $n\geqslant 2$, 令 $\bar{X}=\sum_{i=1}^n X_i/n, S^2=\sum_{i=1}^n(X_i-\bar{X})^2/(n-1)$, 则对任一 $\theta=(\mu,\sigma^2)$, 在 P_θ 之下,

$$\sqrt{n}(\bar{X}-\mu)/S\sim t_{n-1}.$$

由此得出 μ 的 $1-\alpha$ 置信上、下限分别为 $\bar{X}+t_{n-1}(\alpha)S/\sqrt{n}$ 和 $\bar{X}-t_{n-1}(\alpha)S/\sqrt{n}$, 其中 $t_{n-1}(\alpha)$ 为分布 t_{n-1} 的上 α 分位数. 对任何 $\alpha_1>0,\alpha_2>0$, $\alpha_1+\alpha_2=\alpha$, $[\bar{X}-t_{n-1}(\alpha_1)S/\sqrt{n},\bar{X}+t_{n-1}(\alpha_2)S/\sqrt{n}]$ 是 μ 的 $1-\alpha$ 置信区间. 在所有这些置信区间中, 当 $\alpha_1=\alpha_2=\alpha/2$ 时, 该置信区间的长度最小, 此时得到的即为一样本 t-区间估计 $\bar{X}\pm t_{n-1}(\alpha/2)S/\sqrt{n}$.

利用 $(n-1)S^2/\sigma^2\sim\chi^2_{n-1}$, 可得 σ^2 的 $1-\alpha$ 置信上、下限和置信区间分别为

$$(n-1)S^2/\chi^2_{n-1}(1-\alpha),\quad (n-1)S^2/\chi^2_{n-1}(\alpha)$$

和

$$[(n-1)S^2/\chi^2_{n-1}(\alpha_1),(n-1)S^2/\chi^2_{n-1}(1-\alpha_2)],$$

其中 $\alpha_1>0,\alpha_2>0,\alpha_1+\alpha_2=\alpha$. 这种形式的置信区间具有 $[aS^2,bS^2]$ 的形式, 其中 $0<a<b<\infty$, a,b 满足约束条件

$$\int_{(n-1)/b}^{(n-1)/a}k_{n-1}(x)dx=1-\alpha,\tag{4.2}$$

k_{n-1} 为 χ^2_{n-1} 变量的密度函数. 由 Lagrange 乘子法, 若要在这类区间中, 与 a,b 对应的区间长度达到最小, 必须而且只需 a,b 满足 (4.2) 及以下的方程:

$$a^{-2}k_{n-1}\big((n-1)/a\big)=b^{-2}k_{n-1}\big((n-1)/b\big).\tag{4.3}$$

由于该方程的解析解不容易求得, 一般我们常取 $\alpha_1=\alpha_2=\alpha/2$, 所得置信区间与上述区间中最短者相差无几.

例 4.2 设样本 $X_1,\cdots,X_n$ i.i.d., $X_1\sim U(0,\theta)$, $\theta>0$, 要求构建建 θ 置信区间.

记 $X_{(n)}=\max_{1\leqslant i\leqslant n}X_i$, 则 $X_{(n)}/\theta$ 的密度函数为 $nx^{n-1}I(0<x<1)$, 枢轴变量可取为 $X_{(n)}/\theta$. 对给定 $0<\alpha<1$, 找 $0<a<b<1$ 使 $b^n-a^n=1-\alpha$, 则 $[X_{(n)}/b,\ X_{(n)}/a]$ 是 θ 的 $1-\alpha$ 置信区间. 在这些置信区间中, 以取 $b=1,a=\alpha^{1/n}$ 时长度最小. 我们把这一结论的证明留给读者.

例 4.3 设样本 $X_1,\cdots,X_m$ i.i.d., $Y_1,\cdots,Y_n$ i.i.d., 且全体独立, $X_1\sim N(\mu_1,\sigma^2)$, $Y_1\sim N(\mu_2,\sigma^2)$, 其中 $-\infty<\mu_1,\mu_2<\infty,\sigma^2>0$, $m\geqslant 2,n\geqslant 2$. 为了构建 $\delta=\mu_2-\mu_1$ 的置信区间和置信限, 记

$$S^2=\frac{1}{m+n-2}\Big[\sum_{i=1}^m(X_i-\bar{X})^2+\sum_{j=1}^n(Y_j-\bar{Y})^2\Big],$$

则

$$\sqrt{\frac{mn}{m+n}}\big(\bar{Y}-\bar{X}-\delta\big)/S\sim t_{n+m-2}.$$

利用此分布, 可得 δ 的 $1-\alpha$ 置信上限

$$(\bar{Y}-\bar{X})+\sqrt{\frac{m+n}{mn}}t_{n+m-2}(\alpha)S,$$

置信下限

$$(\bar{Y}-\bar{X})-\sqrt{\frac{m+n}{mn}}t_{n+m-2}(\alpha)S,$$

以及置信区间

$$\left[(\bar{Y}-\bar{X})-\sqrt{\frac{m+n}{mn}}t_{n+m-2}(\alpha/2)S,(\bar{Y}-\bar{X})+\sqrt{\frac{m+n}{mn}}t_{n+m-2}(\alpha/2)S\right],$$

此即常见的两样本 t-区间估计 $(\bar{Y}-\bar{X})\pm\sqrt{\frac{m+n}{mn}}t_{n+m-2}(\alpha/2)S$.

下面我们去掉 $\sigma_1^2=\sigma_2^2$ 的假定, 重新考虑有关参数的置信区间的构建问题. 记

$$S_X^2=\frac{1}{m-1}\sum_{i=1}^{m}(X_i-\bar{X})^2,\quad S_Y^2=\frac{1}{n-1}\sum_{i=1}^{n}(Y_i-\bar{Y})^2,$$

则 σ_2^2/σ_1^2 的 $1-\alpha$ 置信限和置信区间, 可利用

$$(S_X^2/\sigma_1^2)/(S_Y^2/\sigma_2^2)\sim F_{m-1,n-1}$$

构建. 但 $\delta=\mu_2-\mu_1$ 的区间估计就不那么简单. 这个问题就是统计学上著名的 **Behrens-Fisher 问题**. 初看起来, 利用

$$(\bar{Y}-\bar{X}-\delta)/\sqrt{\sigma_1^2/m+\sigma_2^2/n}\sim N(0,1),$$

并将分母中的 $\sigma_1^2/m+\sigma_2^2/n$ 用它的无偏估计 $S_X^2/m+S_Y^2/n$ 代入, 似可处理这一问题, 但由此得到的 $T=(\bar{Y}-\bar{X}-\delta)/\sqrt{S_X^2/m+S_Y^2/n}$ 的分布与 (σ_1^2,σ_2^2) 有关, 所以不能像等方差情形一样构建 δ 的区间估计. 在 4.5 节我们将介绍 Fisher 从另一观点提出的一种解法. 基于 T 的渐近正态性, 当 m,n 很大时, 此问题也可用大样本方法来处理.

4.2.2 基于连续随机变量构建置信区间

由 4.2.1 节可知, 为要构建 θ 的 $1-\alpha$ 置信区间, 首先要找出 θ 的一个良好估计量 $T=T(X)$, 并据此寻找枢轴变量, 再构建 θ 的置信区间, 或 θ 的大样本区间估计.

当统计量 $T(X)$ 是连续随机变量时, 寻找枢轴变量不是一件难事. 设 $T(X)$ 的分布函数 $P_\theta\{T(X)\leqslant t\}=G(t,\theta)$, 则 $G(T(X),\theta)$ 服从区间 (0, 1) 上的均匀分布, 此时 $G(T(X),\theta)$ 即可取为枢轴变量.

例 4.4 设 $X_1,\cdots,X_n$ 为取自分布为

$$f(x;\theta)dx=(\theta/x^2)I(\theta<x<\infty)dx$$

的总体的 i.i.d. 样本, 其中参数 $\theta>0$ 未知. 试构建 θ 的 $1-\alpha$ 置信区间.

为此, 注意到最小次序统计量 $X_{(1)}=\min\{X_i:1\leqslant i\leqslant n\}$ 是 θ 的一个充分统计量, 其密度函数和分布函数分别为

$$g(t,\theta)=n\theta^n/t^{n+1}I(t>\theta),\quad 0<\theta<\infty$$

和

$$G(t,\theta)=(1-\theta^n/t^n)I(t>\theta),\quad 0<\theta<\infty.$$

枢轴变量可取为

$$G(X_{(1)},\theta)=1-\theta^n/X_{(1)}^n\sim U(0,1).$$

选取两个常数 c 和 d $(0<c<d<1)$, 满足 $d-c=1-\alpha$, 则有

$$P_\theta(c\leqslant 1-\theta^n/X_{(1)}^n\leqslant d)=1-\alpha,$$

即得 θ 的 $1-\alpha$ 置信区间 $[(1-d)^{1/n}X_{(1)},(1-c)^{1/n}X_{(1)}]$. 不难证明, 若取 $c=0,d=1-\alpha$, 所得置信区间 $[\alpha^{1/n}X_{(1)},X_{(1)}]$ 在上述区间中长度最小.

利用统计量 $T(X)$ 的分布函数 $G(t,\theta)$ 求取 θ 的置信区间时, 需要把不等式 $c\leqslant G(T(X),\theta)\leqslant d$ 等价转换为 $A(X)\leqslant\theta\leqslant B(X)$. 此时, 要求对每个给定的 t, 函数 $G_t(\cdot):=G(t,\cdot)$ 当其取值于 (0, 1) 开区间时严格单调, 因而存在严格单调的反函数 G_t^{-1}. 这个条件在样本 X 的密度族关于 $T(x)$ 具有单调似然比时, 总是可以满足的. 事实上, 若在定理 3.3 的 (i) 中取 $\phi(X)=I(T>t)$, 则对给定的 t, 在集合 $\{\theta:0<1-G(t,\theta)\equiv\mathrm{E}_\theta\phi(X)<1\}$ 上, $1-G(t,\theta)$ 对 θ 严格增. 这一结论也可由第三章习题 3.1 证得.

最后我们指出, 在 4.2.1 和 4.2.2 节的例子中构建的置信集, 其置信系数均为 $1-\alpha$.

4.2.3 基于离散随机变量构建置信区间

设 $G(t,\theta)=P_\theta(T\leqslant t)$ 为 $T(X)$ 的分布函数. 当统计量 $T(X)$ 是离散变量时, $G(T(X),\theta)$ 不服从 $(0,1)$ 区间上的均匀分布. 此时, 为了通过 $G(T(X),\theta)$ 构建 θ 的置信区间, 我们需要下述引理.

引理 4.1 *设 $F(x)$ 是随机变量 X 的分布函数, 则对任意的 $0\leqslant t\leqslant 1$, 有*

$$P\{F(X)\leqslant t\}\leqslant t\leqslant P\{F(X-0)<t\}.$$

证明可参看第五章推论 5.40. 此引理的一个直接的推论即为: 若 X 的分布函数 F 连续, 则 $F(X)\sim U(0,1)$.

关于 θ 的置信区间和置信限, 我们有下述定理.

定理 4.2 *设对每个 t, 函数 $G(t,\theta)$ 在集合 $\{\theta:0<G(t,\theta)<1\}$ 上对 θ 严格减, 则对给定的 α $(0<\alpha<1)$,*

$$\underline{\theta}=\sup_{\theta\in\Theta}\{\theta:\ G(T-0,\theta)\geqslant 1-\alpha\},\tag{4.4}$$

$$\overline{\theta}=\inf_{\theta\in\Theta}\{\theta:\ G(T,\theta)\leqslant\alpha\}\tag{4.5}$$

分别是 θ 的 $1-\alpha$ 置信下限和置信上限. 而 θ 的 $1-\alpha$ 置信区间可取为 $[\hat{\theta}_1, \hat{\theta}_2]$,

$$\hat{\theta}_1 = \sup_{\theta\in\Theta}\{\theta:\ G(T-0,\theta) \geqslant 1-\alpha_1\},\quad \hat{\theta}_2 = \inf_{\theta\in\Theta}\{\theta : G(T,\theta) \leqslant \alpha_2\},$$

其中 $\alpha_1 \geqslant 0, \alpha_2 \geqslant 0, \alpha_1+\alpha_2 = \alpha$. 约定: $\sup\emptyset = \inf\Theta, \inf\emptyset = \sup\Theta$.

若进一步假定 $G(t,\theta)$ 对 θ 连续, 则问题可作如下简化. 此时, 若方程

$$G(T-0,\theta) = 1-\alpha, \tag{4.6}$$

或方程

$$G(T,\theta) = \alpha \tag{4.7}$$

在 Θ 中有解, 则 $\underline{\theta}$ 和 $\overline{\theta}$ 可分别取为方程 (4.6) 和 (4.7) 的唯一解. 当方程无解时, 集合 $\{G(T-0,\ \theta) \geqslant 1-\alpha\}$ (或集合 $\{\theta : G(T,\theta) \leqslant \alpha\}$) 为 $\emptyset$ 或 Θ, 据 (4.4) 和 (4.5) 以及上述约定, $\underline{\theta}$ 取值 $\sup\emptyset = \inf\Theta$ 或 $\sup\Theta$ ($\overline{\theta}$ 取值 $\inf\emptyset = \sup\Theta$ 或 $\inf\Theta$).

证明 由 $G(t,\theta)$ 对 θ 严格减的假定和 $\underline{\theta}$ 的定义, 当 $G(T-0,\theta) < 1-\alpha$ 时, $\theta \geqslant \underline{\theta}$. 由引理 4.1, 对 $\forall\ \theta\in\Theta$, $P_\theta\{\theta \geqslant \underline{\theta}\} \geqslant P_\theta\{G(T-0,\theta) < 1-\alpha\} \geqslant 1-\alpha$, 故 $\underline{\theta}$ 是 θ 的 $1-\alpha$ 置信下限. 类似地, 在 $G(T,\ \theta) > \alpha$ 时, $\theta \leqslant \overline{\theta}$. 据引理 4.1, 对 $\forall\ \theta$, $P_\theta\{\theta \leqslant \overline{\theta}\} \geqslant P_\theta\{G(T,\theta) > \alpha\} = 1 - P_\theta\{G(T,\theta) \leqslant \alpha\} \geqslant 1-\alpha$, 所以 $\overline{\theta}$ 是 θ 的 $1-\alpha$ 置信上限. 对于置信区间, 结论类似可得. ■

也许有人要问, 方程 $G(T-0,\theta) = 1-\alpha$ 无解的情况真会出现吗? 从下面的例子可以看出, 即使 T 为连续随机变量, 也会出现上述方程无解的情况.

例 4.5 设

$$T \sim N(\theta, 1),\quad \Theta = \{\theta : |\theta| \leqslant 1\},$$

则 $G(t,\theta) = \Phi(t-\theta)$ 为 θ 的严格减的连续函数, 当 $t > u_\alpha + 1$ 时, 对所有 $\theta\in\Theta$ 总有 $G(t,\theta) > 1-\alpha$; 当 $t < u_\alpha - 1$ 时, 对所有 $\theta\in\Theta$ 总有 $G(t,\theta) < 1-\alpha$. 因而当给定的 $T > u_\alpha + 1$ 或 $T < u_\alpha - 1$ 时方程 $G(T,\theta) = 1-\alpha$ 皆无解. 前者蕴含 $\{G(T,\theta) \geqslant 1-\alpha\} = \Theta$, $\underline{\theta} = \sup\Theta = 1$; 后者蕴含 $\{G(T,\ \theta) \geqslant 1-\alpha\} = \emptyset$, $\underline{\theta} = \sup\emptyset = \inf\Theta = -1$.

若 $G(t,\theta)$ 关于 θ 严格增, 则它关于 $(-\theta)$ 严格减. 由定理 4.2 可得下述定理.

定理 4.3 设对每个 t, 函数 $G(t,\theta)$ 在集合 $\{\theta : 0 < G(t,\ \theta) < 1\}$ 上对 θ 严格增, 则对给定的 α $(0<\alpha<1)$,

$$\underline{\theta} = \sup_{\theta\in\Theta}\{\theta : G(T,\theta) \leqslant \alpha\}, \tag{4.8}$$

$$\overline{\theta} = \inf_{\theta\in\Theta}\{\theta : G(T-0,\theta) \geqslant 1-\alpha\} \tag{4.9}$$

分别是 θ 的 $1-\alpha$ 置信下限和置信上限. 而 θ 的 $1-\alpha$ 置信区间可取为 $[\hat{\theta}_1, \hat{\theta}_2]$,

$$\hat{\theta}_1 = \sup_{\theta\in\Theta}\{\theta : G(T,\theta) \leqslant \alpha_2\},\quad \hat{\theta}_2 = \inf_{\theta\in\Theta}\{\theta : G(T-0,\theta) \geqslant 1-\alpha_1\},$$

其中 $\alpha_1 \geqslant 0, \alpha_2 \geqslant 0, \alpha_1+\alpha_2 = \alpha$.

进一步, 假定 $G(t,\theta)$ 对 θ 连续. 此时, 若方程

$$G(T,\theta) = \alpha, \tag{4.10}$$

或方程

$$G(T-0,\theta)=1-\alpha \tag{4.11}$$

在 Θ 中有解, 则 $\underline{\theta}$ 和 $\overline{\theta}$ 可分别取为方程 (4.10) 和 (4.11) 的唯一解. 若方程无解, 集合 $\{\theta: G(T,\theta)\leqslant\alpha\}$ (或集合 $\{\theta: G(T-0,\theta)\geqslant 1-\alpha\}$) 为 $\emptyset$ 或 Θ, 据 (4.8) 和 (4.9) 以及上述约定, $\underline{\theta}$ 取值 $\sup\emptyset=\inf\Theta$ 或 $\sup\Theta$ ($\overline{\theta}$ 取值 $\inf\emptyset=\sup\Theta$ 或 $\inf\Theta$).

下面给出一个关于统计量 $T(X)$ 是离散随机变量的例子.

例 4.6 设样本 $X_1,\cdots,X_n$ i.i.d., 服从两点分布 $B(1,p)$, 其中 $0\leqslant p\leqslant 1$. 问题是如何构建 p 的置信区间.

易知, 充分统计量 $T=\sum_{i=1}^n X_i$ 的分布函数 $P_p\{T\leqslant t\}$ 为

$$G(t,p)=\sum_{i=0}^{[t]}\binom{n}{i}p^i(1-p)^{n-i}=\frac{\Gamma(n+1)}{\Gamma([t]+1)\Gamma(n-[t])}\int_p^1 u^{[t]}(1-u)^{n-[t]-1}du,$$

其中 $[t]$ 表示 t $(0\leqslant t\leqslant n)$ 的整数部分. $G(t,p)$ 是 p 的连续的严格减函数.

统计量 T 的值取正整数, 记为 k. 当 $k>0$ 时, 由定理 4.2, p 的 $1-\alpha$ 置信下限是方程

$$\sum_{i=0}^{k-1}\binom{n}{i}p^i(1-p)^{n-i}=1-\alpha$$

的解, 即方程

$$Be(p\,|k,n-k+1)\equiv\frac{\Gamma(n+1)}{\Gamma(k)\Gamma(n-k+1)}\int_0^p u^{k-1}(1-u)^{n-k}du=\alpha \tag{4.12}$$

的解, 其中 $Be(\cdot|m,n)$ 表示 Beta 分布 $Be(m,n)$ 的分布函数. 不难验证, 若 $B\sim Be(m,n)$, 则 $F:=(n/m)[B/(1-B)]\sim F(2m,2n)$, 故方程 (4.12) 等价于

$$F\Big(\frac{p}{1-p}\frac{n-k+1}{k}\;\Big|\;2k,2(n-k+1)\Big)=\alpha.$$

为了方便, 此处把自由度为 m,n 的中心 F 分布记为 $F(m,n)$, 将其分布函数记为 $F(\cdot|m,n)$, 将它的上 α 分位数记为 $F_\alpha(m,n)$. 因此 p 的 $1-\alpha$ 置信下限 $\underline{p}$ 可取为下列方程的解:

$$\frac{p}{1-p}\frac{n-k+1}{k}=F_{1-\alpha}(2k,2(n-k+1)).$$

由于 $F_{1-\alpha}(m,n)=1/F_\alpha(n,m)$, 所以

$$\underline{p}=\frac{k}{k+(n-k+1)F_\alpha(2(n-k+1),2k)},$$

当 $k=0$ 时, 取 $\underline{p}=0$.

类似地, 由定理 4.2, p 的 $1-\alpha$ 置信上限 $\overline{p}$ 可取为方程

$$\sum_{i=0}^{k}\binom{n}{i}p^i(1-p)^{n-i}=\alpha$$

的解, 即方程

$$\frac{p}{1-p}\frac{n-k}{k+1}=F_\alpha(2k+2,2n-2k)$$

的解, 所以

$$\overline{p}=\frac{(k+1)F_\alpha(2k+2,2n-2k)}{(n-k)+(k+1)F_\alpha(2k+2,2n-2k)},$$

当 $k=n$ 时, 取 $\overline{p}=1$.

由上述结果可知, 在 $0<k<n$ 时, p 的 $1-\alpha$ 置信区间可取为 $[\hat{\theta}_1,\hat{\theta}_2]$, 其中

$$\hat{\theta}_1=\frac{k}{k+(n-k+1)F_{\alpha_1}(2(n-k+1),2k)},$$
$$\hat{\theta}_2=\frac{(k+1)F_{\alpha_2}(2k+2,2n-2k)}{(n-k)+(k+1)F_{\alpha_2}(2k+2,2n-2k)},$$

此处, $\alpha_1\geqslant 0,\alpha_2\geqslant 0$, 且 $\alpha_1+\alpha_2=\alpha$. 在实际应用中, 常取 $\alpha_1=\alpha_2=\alpha/2$.

当 $k=0$ 时, 取

$$\hat{\theta}_1=0,\quad \hat{\theta}_2=\frac{F_\alpha(2,2n)}{n+F_\alpha(2,2n)}.$$

当 $k=n$ 时, 取

$$\hat{\theta}_1=\frac{n}{n+F_\alpha(2,2n)},\quad \hat{\theta}_2=1.$$

4.2.4 假设检验法

这是构建区间估计的另一种常用方法. 为了构建参数 θ 的一个置信水平为 $1-\alpha$ 的置信集, 对每个 $\theta_0\in\Theta$, 考虑检验问题 $\theta=\theta_0\leftrightarrow\theta\in K(\theta_0)$, 其中 $K(\theta_0)$ 是适当选取的 Θ 的一个不含 θ_0 的子集. 设 $A(\theta_0)$ 是此检验问题的一个水平为 α 的接受域, 定义集合

$$S(x)=\{\theta\in\Theta:\ x\in A(\theta)\}$$

(参看图 4.1), 则因

$$P_{\theta_0}(\theta_0\in S(X))=P_{\theta_0}(X\in A(\theta_0))\geqslant 1-\alpha,\quad \text{对任一 } \theta_0\in\Theta,$$

$S(x)$ 是 θ 的一个置信水平 $1-\alpha$ 的置信集. 反之, 若 $S(x)$ 是 θ 的一个置信水平为 $1-\alpha$ 的置信集, 则 $A(\theta_0)=\{x:\ \theta_0\in S(x)\}$ 是检验问题 $\theta=\theta_0\leftrightarrow\theta\in K(\theta_0)$ 的一个水平为 α 的接受域.

用此法构建区间估计的具体做法, 可通过下面的例子来说明.

例 4.7 设 $X_1,\cdots,X_n$ 为取自正态总体 $\{N(\mu,\sigma^2),-\infty<\mu<\infty,\sigma^2>0\}$ 的 i.i.d. 样本. 为构建 μ 的置信区间, 任取 $\mu_0\in\mathbf{R}$, 考虑双边检验问题 $\mu=\mu_0\leftrightarrow\mu\neq\mu_0$. 水平为 α 的双边 t 检验的接受域为

$$|\bar{X}-\mu_0|\leqslant t_{n-1}(\alpha/2)S/\sqrt{n},$$

图 4.1

其中, $\bar{X}=\sum_{i=1}^{n}X_i/n, S^2=\sum_{i=1}^{n}(\bar{X}-X_i)^2/(n-1)$. 将上述接受域写成如下的等价形式:

$$\bar{X}-t_{n-1}(\alpha/2)S/\sqrt{n}\leqslant\mu_0\leqslant\bar{X}+t_{n-1}(\alpha/2)S/\sqrt{n},$$

即得 μ 的一个置信水平为 $1-\alpha$ 的置信区间 $[\bar{X}-t_{n-1}(\alpha/2)S/\sqrt{n},\bar{X}+t_{n-1}(\alpha/2)S/\sqrt{n}]$, 此即常用的 t-区间估计.

如要得到相应的置信上限, 则可考虑单边检验问题 $\mu=\mu_0\leftrightarrow\mu<\mu_0$. 水平为 α 的一个接受域为 $\bar{X}-\mu_0\geqslant-t_{n-1}(\alpha)S/\sqrt{n}$, 由此可导出 μ 的一个置信水平为 $1-\alpha$ 的置信上限 $\bar{X}+t_{n-1}(\alpha)S/\sqrt{n}$.

读者不难用这种方法得出例 4.1 至例 4.3 中其余所有的区间估计. 这种方法的一个主要优点是, 它把一个置信区间 (区域) 与一个检验对应起来, 而检验的优良性质可转化为置信区间的优良性质. 这一方面的有关理论将在后面讨论.

4.2.5 大样本方法

要使区间估计的置信系数严格等于指定的值, 需要知道有关统计量的确切分布. 除了少数情况, 这很难做到. 但若知道当样本量 $n\to\infty$ 时有关统计量的渐近分布, 则在样本量 n 较大时, 可用此渐近分布代替确切分布构建有关参数的置信区间或置信限. 这种方法就是大样本方法. 设相应的区间估计为 J_n, 如果

$$\liminf_{n\to\infty}P\{\theta\in J_n(X)\}\geqslant1-\alpha,\quad\text{对每个 }\theta\in\Theta,$$

则称 $\{J_n\}$ (或 J_n) 有**渐近置信系数** $1-\alpha$. 但因上述收敛对 $\theta\in\Theta$ 一般并无一致性, 在上式成立时, 对给定的 n, J_n 的置信系数往往低于 $1-\alpha$.

例 4.8 将上述方法用于二项变量 $X_n\sim B(n,\theta),0<\theta<1$, 记 $\hat{\theta}_n=\bar{X}$, 由

$$\sqrt{n}(\hat{\theta}_n-\theta)/\sqrt{\theta(1-\theta)}\xrightarrow{L}N(0,1),$$

有 $\lim_{n\to\infty}P_\theta\{\sqrt{n}|\hat{\theta}_n-\theta|/\sqrt{\theta(1-\theta)}\leqslant u_{\alpha/2})\}=1-\alpha$, 解出 θ 即得 θ 的区间估计

$$J_n^*:=\frac{n}{n+c^2}\Big(\hat{\theta}_n+\frac{c^2}{2n}\pm c\sqrt{\hat{\theta}_n(1-\hat{\theta}_n)/n+c^2/4n^2}\Big),\quad c=u_{\alpha/2},$$

其渐近置信系数为 $1-\alpha$.

若应用极限分布

$$\sqrt{n}(\hat{\theta}_n-\theta)/\sqrt{\hat{\theta}_n(1-\hat{\theta}_n)}\xrightarrow{L}N(0,1),$$

即得 θ 的另一个大样本区间估计

$$J_n:=\hat{\theta}_n\pm u_{\alpha/2}\sqrt{\hat{\theta}_n(1-\hat{\theta}_n)/n},$$

其渐近置信系数也是 $1-\alpha$.

不难证明, 当 $n\geqslant 2$ 给定时, J_n 的置信系数必为 0. 事实上, 当 $X_n=n$ 即 $\hat{\theta}_n=1$ 时, 由上式, J_n 退化为单点集 $\{1\}$, 因 $\theta<1$, J_n 不包含 θ, 故

$$P_\theta\{\theta\in J_n\}\leqslant 1-P_\theta\{X_n=n\}=1-\theta^n,$$

因其在参数空间 $(0,1)$ 区间的下确界为 0, 按定义, J_n 的置信系数为 0.

区间估计 J_n 由于形式简单而受到欢迎, 但由于用到了正态逼近和标准差估计两次逼近, 使用时样本量应更大一些.

在实际应用中, 为达到区间估计的某种精度要求, 常常要估计所需要的样本量. 大样本方法可用来近似地作出此项估计. 例如, 某市要了解公众对某项政策的支持率 θ, 问需要随机地调查多少居民, 方能使估计值 $\hat{\theta}_n$ 和真值 θ 间的最大误差不超过 0.03, 且置信系数为 95%? 取 $u_{0.025}=1.96$, 则当采用 J_n 作区间估计时, 要求 $u_{0.025}[\hat{\theta}_n(1-\hat{\theta}_n)/n]^{1/2}\leqslant 0.03$, 由于 $\hat{\theta}_n(1-\hat{\theta}_n)$ 最大可能达到 0.25, 故样本量要满足近似的不等式 $n\geqslant(1.96\times 0.5/0.03)^2\approx 1067$.

一般地, 设 $\mathcal{F}$ 为方差非 0 有限的某个一维分布 F 的分布族, $X_1,\cdots,X_n$ 为从中抽取的 i.i.d. 样本, 要作分布泛函 $\theta=\theta(F)=\int_{-\infty}^{\infty}xdF$ 的区间估计. 记 $\bar{X}_n=\sum_{i=1}^n X_i/n, S_n^2=\sum_{i=1}^n(\bar{X}-X_i)^2/n$, 则

$$\sqrt{n}(\bar{X}_n-\theta)/S_n\rightsquigarrow N(0,1).$$

近似地把 $N(0,1)$ 作为 $\sqrt{n}(\bar{X}_n-\theta)/S_n$ 的分布, 即得 θ 的渐近置信系数 $1-\alpha$ 的区间估计 $J_n=\bar{X}_n\pm u_{\alpha/2}S_n/\sqrt{n}$.

在一些正则条件下, 参数 θ 的 MLE $\hat{\theta}_n$ 是渐近有效的, $\sqrt{n}(\hat{\theta}_n-\theta)\xrightarrow{L}N(0,\sigma^2(\theta))$, 当 $\sigma(\cdot)$ 连续时即有 $\sqrt{n}(\hat{\theta}_n-\theta)/\sigma(\hat{\theta}_n)\xrightarrow{L}N(0,1)$. 由此可导出 θ 的区间估计 $\hat{\theta}_n\pm u_{\alpha/2}\sigma(\hat{\theta}_n)/\sqrt{n}$, 渐近置信系数为 $1-\alpha$. 这种方法也适用于多维参数的区域估计. 与所有大样本区间估计一样, 其共同的缺点在于, 对固定的 n, 很难知道真实的置信系数.

4.3 区间估计的优良性

所谓集估计或区间估计的优良性, 是指在一定置信水平下, 精度好者为优.

设对任一 $\theta\in\Theta$, W_θ 是 Θ 的一个不包含 θ 的子集. 未知参数 θ 的一个置信水平为 $1-\alpha$ 的置信集 $J^*(X)$ 称为 **W_θ-一致最精确的** (Uniformly Most Accurate, UMA), 如果对于置信水平为 $1-\alpha$ 的任一置信集 $J(X)$ 和任一 $\theta_0\in\Theta$, 有

$$P_{\theta_0}\big(\theta\in J^*(X)\big)\leqslant P_{\theta_0}(\theta\in J(X)),\ \theta\in W_{\theta_0}. \tag{4.13}$$

在上述定义中, 若取 $W_\theta = \{\theta\}^c \equiv \Theta \setminus \{\theta\}$, 则称 J^* 是 θ 的置信水平为 $1-\alpha$ 的**一致最精确的置信集**, 简称 $1-\alpha$ **UMA 置信集 (区间)**.

若考虑的是一维参数 θ 的置信上限(或置信下限), 通常在定义中只需取 $W_\theta = (\theta, \infty)$ (或 $(-\infty, \theta)$), 此时称 J^* 为 $1-\alpha$ **UMA 置信上限(或置信下限)**.

若在上述 UMA 置信集的定义中, 所考察的仅限于指定的某个类 $\mathcal{F}$ 中的置信集, 则可在相应的 $1-\alpha$ UMA 置信集前冠以 "$\mathcal{F}$ 类中的" 字样.

称 J 是 θ 的一个 $1-\alpha$ **无偏置信集**, 如果

$$P_{\theta_0}(\theta_0 \in J) \geqslant 1-\alpha \geqslant P_{\theta_0}(\theta \in J), \quad \text{对任何 } \theta_0 \in \Theta,\ \theta \in \Theta, \theta \neq \theta_0. \tag{4.14}$$

在 $1-\alpha$ UMA 置信集 的定义中, 若将 $\mathcal{F}$ 取成 θ 的全体 $1-\alpha$ 无偏置信集的类, 则当 (4.13) 成立时, 称 J^* 为 θ 的 $1-\alpha$ **一致最精确的无偏置信集**, 简称为 $1-\alpha$ **UMAU 置信集**.

对无偏检验我们曾有过类似的定义: 称 ϕ 是检验问题 "$H \leftrightarrow K$" 的一个水平为 α 的无偏检验, 如果

$$\beta_\phi(\theta) \leqslant \alpha \leqslant \beta_\phi(\tilde{\theta}), \quad \text{对任何 } \theta \in H, \tilde{\theta} \in K.$$

在上述定义之下, 通过 $x \in A(\theta_0) \Leftrightarrow \theta_0 \in S(x)$ 的等价性, 在 "$\theta = \theta_0 \leftrightarrow \theta \neq \theta_0$" 的以 $A(\theta_0)$ 为接受域的检验 ϕ 和 θ 的置信集 $S(X)$ 之间确定了如下的关系: **若 ϕ 为水平 α 无偏, 则 $S(X)$ 为 $1-\alpha$ 无偏, 反之亦然.** 事实上, 若 ϕ 为水平 α 无偏, 则按定义, 对任何 $\theta \neq \theta_0$, $P_\theta\{X \in A(\theta_0)\} \leqslant 1-\alpha$, 即 $P_\theta\{\theta_0 \in S(X)\} \leqslant 1-\alpha$. 而

$$P_{\theta_0}\{\theta_0 \in S(X)\} = P_{\theta_0}\{X \in A(\theta_0)\} \geqslant 1-\alpha,$$

上面两式可合并写成

$$P_{\theta_0}\{\theta_0 \in S(X)\} \geqslant 1-\alpha \geqslant P_{\theta_0}\{\theta \in S(X)\}, \quad \text{对任何 } \theta \neq \theta_0,$$

证得 $S(X)$ 为 θ 的 $1-\alpha$ 无偏置信集. 反之, 若 $S(X)$ 为 $1-\alpha$ 无偏, 则上式成立, 即

$$P_{\theta_0}\{X \in A(\theta)\} \leqslant 1-\alpha \leqslant P_{\theta_0}\{X \in A(\theta_0)\}, \quad \text{对任何 } \theta \neq \theta_0,$$

上述不等式可改写为

$$P_{\theta_0}\{X \notin A(\theta_0)\} \leqslant \alpha \leqslant P_\theta\{X \notin A(\theta_0)\}, \quad \text{对任何 } \theta \neq \theta_0,$$

即证明了以 $A(\theta_0)$ 为接受域的检验, 是 "$\theta = \theta_0 \leftrightarrow \theta \neq \theta_0$" 的一个水平为 α 的无偏检验.

对置信上 (下) 限, 上述检验问题要改为 "$\theta = \theta_0 \leftrightarrow \theta < \theta_0$" (相应地, "$\theta = \theta_0 \leftrightarrow \theta > \theta_0$"). 这样, 区间估计和假设检验无偏性的这两个定义在两者的优良性之间就架起了一座桥梁.

有了以上的准备, 我们就可以将置信集的优良性定理, 写成下面统一的形式.

定理 4.4 沿用 4.2.4 节开头的符号, 设由置信集 $S(X)$ 对应的检验 ϕ, 对任何 θ_0, 是 "$\theta = \theta_0 \leftrightarrow \theta \in K(\theta_0)$" 的水平为 α 的 UMP 检验, 则 $S(X)$ 是 $1-\alpha$ UMA 置信集. 上述结论自然适用于一维参数 θ 的置信区间和置信上 (下) 限, 在置信区间的情形, 相应的检验问题取为形如 "$\theta = \theta_0 \leftrightarrow \theta \neq \theta_0$" 的双边检验, 而在置信上限或置信下限情形, 相应的检验问题取为形如 "$\theta = \theta_0 \leftrightarrow \theta < \theta_0$" 或 "$\theta = \theta_0 \leftrightarrow \theta > \theta_0$" 的

单边检验.

若加上 "无偏" 两字, 则上述结论和下面的证明一字不改地成立.

证明 以置信上限 J 为例. 设置信上限 J 对应的检验 ϕ, 对任意给定的 θ_0, 是 "$\theta=\theta_0 \leftrightarrow \theta<\theta_0$" 的水平为 α 的 UMP 检验, 其接受域为 $A(\theta_0)$. 由 ϕ 有水平 α 知 J 有置信水平 $1-\alpha$. 今设 $\tilde{J}$ 为另一 $1-\alpha$ 置信上限, 则与其相应的检验 $\tilde{\phi}$, 即为同一检验问题的一个水平为 α 的检验, 其接受域记为 $\tilde{A}(\theta_0)$. 因 ϕ 为 UMP, 故对任一 $\theta<\theta_0$ 应有 $\beta_{\tilde{\phi}}(\theta)\leqslant\beta_{\phi}(\theta)$, 即 $P_\theta\{X\in A(\theta_0)\}\leqslant P_\theta\{X\in\tilde{A}(\theta_0)\}$. 因 $\theta_0\in(-\infty,J(X)]\Leftrightarrow X\in A(\theta_0)$, $\theta_0\in(-\infty,\tilde{J}(X)]\Leftrightarrow X\in\tilde{A}(\theta_0)$, 故

$$P_\theta\{\theta_0\in(-\infty,J(X)]\}\leqslant P_\theta\{\theta_0\in(-\infty,\tilde{J}(X)]\},\quad \text{对任何 } \theta<\theta_0,$$

从而证明了 J 是 $1-\alpha$ UMA 置信上限. ■

上述定理的逆一般不成立. 这是因为, 在定理和证明中涉及的检验都只限于非随机化检验. 基于上述证明过程只能证明, 若 J 为 θ 的 $1-\alpha$ UMA 置信上限, 则相应的检验 ϕ, 对任一 θ_0, 在 "$\theta=\theta_0\leftrightarrow\theta>\theta_0$" 水平 α 的所有**非随机化**检验中为 UMP, 一般不能把**非随机化**这一限制词去掉.

关于单调似然比分布族, 以及指数族中一维参数的 $1-\alpha$ UMA 或 UMAU 置信区间和置信限的详细讨论, 有兴趣的读者可参看陈希孺 (1999, 2009) 的 7.2 节的有关定理. 相应的陈述虽然复杂, 好在这些 UMA 或 UMAU 置信集, 均可通过定理 4.4, 从定理 3.3, 3.9 或 3.10 中对应的水平 α 的 UMP 或 UMPU 检验转换而来. 了解了这一点, 在一些具体的例子中, 有关参数的 UMA 或 UMAU 置信区间和置信限就都不难得到了.

在本章开头, 我们介绍了衡量区间估计精度的两种准则, 一种准则导致上述 $1-\alpha$ UMA 或 UMAU 置信区间, 另一种则导致平均长度最短的 $1-\alpha$ 置信区间. 由于准则不同, 两种最优置信区间也往往是不同的. 下面的定理给出了两者之间的一个联系.

定理 4.5 设参数空间 Θ 是直线上含有内点的一个区间. 如果 $[a(X),\ b(X)]$ 是 θ 的 $1-\alpha$ UMA $\mathcal{F}$ 置信区间, 则对 θ 的任一 $1-\alpha$ 的 $\mathcal{F}$ 置信区间 $[a^*(X),\ b^*(X)]$, 有

$$\mathrm{E}_\theta[b(X)-a(X)]\leqslant\mathrm{E}_\theta[b^*(X)-a^*(X)],\quad \text{对所有 } \theta\in\Theta.$$

证明 由 Fubini 定理,

$$\begin{aligned}\mathrm{E}_\theta[b(X)-a(X)]&=\int_{\mathcal{X}}dP_\theta(x)\int_{a(x)}^{b(x)}d\theta_1=\int_\Theta d\theta_1\int_{\{x:a(x)\leqslant\theta_1\leqslant b(x)\}}dP_\theta(x)\\&=\int_\Theta P_\theta\{a(x)\leqslant\theta_1\leqslant b(x)\}d\theta_1=\int_{\theta_1\neq\theta}P_\theta\{a(x)\leqslant\theta_1\leqslant b(x)\}d\theta_1\\&\leqslant\int_{\theta_1\neq\theta}P_\theta\{a^*(x)\leqslant\theta_1\leqslant b^*(x)\}d\theta_1=\mathrm{E}_\theta[b^*(X)-a^*(X)].\end{aligned}$$

因此, 定理成立. ■

由此定理, $1-\alpha$ UMA 置信区间, 其平均长度在所有 $1-\alpha$ UMA 置信区间的类中最短; $1-\alpha$ UMAU 置信区间, 其平均长度在所有 $1-\alpha$ UMAU 置信区间的类中

最短. 只是对于相应的双边检验 "$\theta=\theta_0 \leftrightarrow \theta\neq\theta_0$" 而言, 存在水平为 α 的 UMP 的情形只是凤毛麟角, 因此定理在 UMA 置信区间方面的应用是极其有限的.

易见, 定理可以类推到 Θ 为多维, ν 为 $\mathcal{B}_\Theta$ 上一般测度的情形, 只需假定 $\nu(\Theta)>0$ 且每个单点集的 ν 测度为 0.

上述定理不适用于置信限, 即使与之等价的置信集形式上也是一个有界区间. 例如在指数分布族 $\{E(0,\theta),\theta>0\}$ 中, θ 的置信上限 $J(X)$ 就是这种情形. 反例见 Madansky (1962), 或者参看陈希孺 (1981, 1997).

由定理 4.4, 结合定理 3.3、定理 3.9 和定理 3.10, 或者结合 3.4.4 节与正态有关的检验, 不难得到以下结论.

例 4.9 设 $X_1,\cdots,X_n$ i.i.d., $X_1\sim N(\mu,\sigma^2)$, 分别以 $\bar X$ 和 S_n^2 记样本均值和样本方差, 则当 $\sigma^2>0$ 已知时, $\bar X\pm u_{\alpha/2}\sigma/\sqrt{n}$ 为 μ 的 $1-\alpha$ UMA 置信区间, $\bar X+u_\alpha\sigma/\sqrt{n}$ 和 $\bar X-u_\alpha\sigma/\sqrt{n}$ 分别为 μ 的 $1-\alpha$ UMA 置信上限和置信下限;

μ 和 σ^2 皆未知时, $\bar X\pm t_{n-1}(\alpha/2)S/\sqrt{n}$, $\bar X+t_{n-1}(\alpha)S/\sqrt{n}$ 和 $\bar X-t_{n-1}(\alpha)S/\sqrt{n}$ 分别为 μ 的 $1-\alpha$ UMAU 置信区间, 置信上限和置信下限. 由 3.4.4 节, 检验问题 "$\sigma^2=\sigma_0^2\leftrightarrow\sigma^2\neq\sigma_0^2$" 的水平为 α 的 UMPU 检验有形如 $c_1\sigma_0^2\leqslant(n-1)S^2\leqslant c_2\sigma_0^2$ 的接受域, c_1,c_2 满足

$$\int_{c_1}^{c_2}k_{n-1}(x)dx=1-\alpha,\quad c_1k_{n-1}(c_1)=c_2k_{n-1}(c_2),\tag{4.15}$$

此处 k_{n-1} 为 χ^2_{n-1} 的密度函数. 故 $[(n-1)S^2/c_2,\ (n-1)S^2/c_1]$ 即为 σ^2 的 $1-\alpha$ UMAU 置信区间, 其中 c_1,c_2 满足 (4.15). 由定理 4.5, 该区间的平均长度在 σ^2 的 $1-\alpha$ 无偏置信区间中为最小; 但在所有水平为 $1-\alpha$ 的置信区间中并不为最小. 参看例 4.1.

设样本 $X_1,\cdots,X_n$ i.i.d., 服从指数分布 $E(0,\theta):\theta>0$. 由定理 4.4, 结合习题 3.4, $2\sum_{i=1}^n X_i/\chi^2_{2n}(1-\alpha)$ 和 $2\sum_{i=1}^n X_i/\chi^2_{2n}(\alpha)$ 分别为 θ 的 $1-\alpha$ UMA 置信上限和置信下限.

由定理 4.4, 结合习题 3.10, 若 $X_1,\cdots,X_n$ i.i.d., 服从均匀分布 $U(0,\theta):\theta>0$, 且记 $M=\max_{1\leqslant i\leqslant n}X_i$, 则 $[M,\alpha^{-1/n}M]$ 为 θ 的 $1-\alpha$ UMA 置信区间. 另外可得, $\alpha^{-1/n}M$ 和 $(1-\alpha)^{-1/n}M$ 分别为 θ 的 $1-\alpha$ UMA 置信上限和置信下限.

下面我们介绍一种简便的方法, 用以在基于枢轴变量的一类 $1-\alpha$ 置信区间中, 寻求长度最小的置信区间. 为此, 我们引进如下术语.

设随机变量 $U\sim dQ(u)=q(u)du$. 如果存在常数 $c>0$, 使

$$J=\{u:q(u)\geqslant c\},\ \text{且}\ \int_J g(u)du=1-\alpha,\tag{4.16}$$

则称 J 是 Q-概率为 $1-\alpha$ 的**最高概率密度区域**, 简记为 **HPD 区域**.

我们有下述定理. 定理的结论是显而易见的因而无需证明.

定理 4.6 *设 θ 是一维未知参数, $T=T(X),V=V(X)$ 都是实值统计量, $V(X)$ 恒正, $(\theta-T)/V$ 为一枢轴变量, 其分布为 $dQ(u)=q(u)du$. 若存在常数 $a^*<b^*$ 使区*

间 $[a^*, b^*]$ 与上述 HPD 区域一致, 则在 θ 的形如 $[T+aV,\ T+bV]$ 的 $1-\alpha$ 置信区间中, $[T+a^*V,\ T+b^*V]$ 的长度为最小.

若取 $T \equiv 0$, 并限制 $0 \leqslant a < b, 0 \leqslant a^* < b^*$, 则上述结论也适用于刻度参数 $\theta > 0$ 的置信区间.

4.4 Bayes 区间估计

前几节中介绍的区间估计, 是目前应用最为广泛的 Neyman 模式. 其特点在于, 它可以完全纳入现代概率论的理论框架之内, 允许有一种频率解释; 在推断方面则完全基于样本分布, 不需要其他附加信息. 将这种方法用于若干常见分布时, 能够给出形式简单且符合置信要求的区间估计. 但为了得到符合要求的确切解, 必须求出有关统计量的精确分布, 这在复杂一些的分布族中, 是一件很难做到的事, 此时不得不借助于近似解法, 例如大样本方法. Bayes 方法利用后验分布来构建区间估计, 则可以弥补这个缺陷.

设样本空间为 $(\mathcal{X}, \mathcal{B})$, 样本分布为 $dP_\theta(x) = f(x,\theta)d\mu(x), x \in \mathcal{X};\ \theta \in \Theta,\ \xi$ 为在 $(\Theta, \mathcal{B}_\Theta)$ 上给定的先验分布, 则在有了样本 $X = x$ 后, θ 的后验分布为

$$P(d\theta \mid x) = f(x,\theta)d\xi(\theta) \Big/ \int_\Theta f(x,\theta)d\xi(\theta). \tag{4.17}$$

关于 θ 的 Bayes 推断即基于这一后验分布. 特别地, 如果集合 $J(x)$ 满足条件

$$P(J(x) \mid x) \geqslant 1-\alpha, \text{ 对任一 } x \in \mathcal{X}, \tag{4.18}$$

则称 $J(x)$ 是 θ 的一个**可信水平** (credible level) $1-\alpha$ 的 **Bayes 可信区域** (Bayes credible region). **如有可能, 总是取** $J(x)$ **使上式中的等号成立**, 此时 $1-\alpha$ 称为 $J(x)$ 的**可信系数**, 这与 Neyman 意义下的置信系数不是一回事. 显然, 可信系数为 $1-\alpha$ 的 Bayes 可信区域常常有许多个. 在后验分布关于 Lebesgue 测度有密度 $\pi(\theta \mid x)$ 的场合, 通常推荐的是满足条件

$$J(x) = \{\theta : \pi(\theta \mid x) \geqslant c(x)\} \ \text{ 且 } \int_{J(x)} \pi(\theta \mid x)d\theta \equiv 1-\alpha \tag{4.19}$$

的 Bayes 可信区域 $J(x)$, 称为可信系数 $1-\alpha$ 的**最高后验密度区域** (highest posterior density region), 简称 **HPD 区域**. HPD 区域的形状由后验密度决定, 有可能会由几个连通部分组成. 尽管 **HPD 区域** 在此处和 4.3 节中有不同的含义, 但一般不会引起混淆.

若一维 Bayes 可信区域具有 $[\hat{\theta}_1(x), \hat{\theta}_2(x)]$ 的形式, 则称为 θ 的一个 **Bayes 可信区间**; 若其具有 $(-\infty, \overline{\theta}(x)]$ (或 $[\underline{\theta}(x), \infty)$) 的形式, 则称 $\overline{\theta}(x)$ (或 $\underline{\theta}(x)$) 为 θ 的 **Bayes 可信上限** (**可信下限**). 在给定的先验分布之下, θ 的可信系数为 $1-\alpha$ 的 Bayes 可信上、下限都是唯一的. 易见, **给定可信系数** $1-\alpha$**, 若 HPD 区域也是一个区间, 则这一区间的后验长度在所有** $1-\alpha$ **Bayes 可信区间中必为最小**.

在有些文献中, 可信系数又称为**后验置信系数**或 **Bayes 置信系数**.

这种区间估计方法的优点, 主要有以下三点.

(1) 假定根据具体样本 x, 求得 θ 的可信系数 0.95 的可信区间为 [1.5, 2.1], 这时我们可以写成

$$P(1.5 \leqslant \theta \leqslant 2.1 \mid x) = 0.95,$$

即给定样本 x, θ 落入这个区间的条件概率为 0.95. 但在经典统计中, θ 只是一个未知常数, 它要么在区间 [1.5, 2.1] 上, 要么在这个区间之外, 不能说 "θ 落在区间 [1.5, 2.1] 的概率为 0.95", 而只能说: 这种区间估计在大量重复使用时, 可以期望它覆盖 θ 的频率接近于 95%. 如果我们只是一次两次用此估计, 这种频率解释是没有什么意义的. 相比之下, Bayes 可信区间简单、自然, 易于为人们理解和接受.

(2) 为了得出 Neyman 区间估计, 需要寻求有关统计量的抽样分布, 这常常是一个难题. 而对 Bayes 区间估计方法而言, 无论是计算后验分布, 还是寻求 $J(x)$ 满足 (4.18), 在数学方面都没有原则困难. 连带的问题是从精度方面比较估计优劣的问题. 如上所述, 给定可信水平的最短的可信区间一般不难确定. 如以长度作为衡量精度的优良性准则, 则这一区间就是在此水平下最优的区间估计. 但在 Neyman 机制下, 只有少数情形可以求得最优解.

(3) 因后验分布不涉及测度 μ, 故在样本分布为离散的情形, 构建 Bayes 可信区间没有什么困难; 但这在 Neyman 架构下却是一件有些棘手的事情.

由于这些优点, Bayes 区间估计方法在应用方面颇受欢迎. 但这种方法也有其本身的问题, 即如何确定先验分布的问题会带来困扰. 在实际应用中, 先验分布经常由历史经验来确定. 关于Bayes区间估计的详细介绍, 可参看 Lindley (1965), Box 和 Tiao (1973), Berger (1985) 和 Rubin (1984).

例 4.10 设样本 $X_1, \cdots, X_n$ i.i.d., 服从正态分布 $N(\theta, \sigma^2)$, 其中 σ^2 已知. 取 θ 的先验分布为 $N(\nu, \gamma^2)$, 要求构建 θ 的 Bayes 可信区间.

由 2.3.2 节例 2.14, 给定样本 $X := (X_1, \cdots, X_n)^\tau$, θ 的后验分布为 $N(\theta_X, \sigma_X^2)$, 其中

$$\theta_X = (n\sigma^{-2}\bar{X} + \gamma^{-2}\nu)/(n\sigma^{-2} + \gamma^{-2}), \quad \sigma_X^2 = 1/(n\sigma^{-2} + \gamma^{-2}),$$

其中 $\bar{X} = (1/n)\sum_{i=1}^n X_i$. 由此可得, 可信系数为 $1-\alpha$ 的最短 Bayes 可信区间为 $[\theta_X - u_{\alpha/2}\sigma_X,\ \theta_X + u_{\alpha/2}\sigma_X]$, Bayes 可信上、下限分别为 $\theta_X + u_\alpha\sigma_X$ 和 $\theta_X - u_\alpha\sigma_X$. 上述 Bayes 可信区间不同于 Neyman 置信区间 $\bar{X} \pm u_{\alpha/2}\sigma/\sqrt{n}$. 后者可视为前者当 $\gamma \to \infty$ 时的极限. 进一步可以证明, 不论取何种狭义先验分布, 可信系数为 $1-\alpha$ 的 Bayes 可信区间都不可能与 Neyman 区间 $\bar{X} \pm u_{\alpha/2}\sigma/\sqrt{n}$ 相同. 参看 陈希孺 (1999, 2009) 第七章习题 48.

与 Bayes 点估计情形一样, 在寻求 Bayes 区间估计时, 也可以使用广义先验分布 ξ, 此时后验分布仍有 (4.17) 的表达式, 由此得出的 Bayes 区间估计称为广义 Bayes

可信区间. 在本例中, 若将先验分布取为无信息先验 $d\xi(\theta)=d\theta$, 则 θ 的后验分布为 $N(\bar{X},\ \sigma^2/n)$, 此时, 可信系数为 $1-\alpha$ 的可信区间和 Neyman 置信系数为 $1-\alpha$ 的置信区间形式上完全一样.

在下面的例子中会涉及不同形式的 HPD 区域.

例 4.11 设样本 X 服从二项分布 $B(n,\theta)$, 取其共轭先验 Beta 分布 $Be(a,b)$, 密度函数为

$$\frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)}\theta^{a-1}(1-\theta)^{b-1}I(0<\theta<1).$$

由 2.3.2 节例 2.13, 给定样本 X, θ 的后验分布即为 $Be(a+x,b+n-x)$. 基于后验密度 $\pi(\theta\mid x)$, 可以构建许多区间 $J(x)$ 满足 $P(J(x)\mid x)\geqslant 1-\alpha$, 但常被推荐的就是满足 (4.19) 的 HPD 区域. 固定 x, 后验密度 $\pi(\theta\mid x)$ 只有以下几种情形:

(a) $\pi(\theta|x)$ 是减函数;

(b) $\pi(\theta|x)$ 是增函数;

(c) 对于某个 θ_0, 函数 $\pi(\theta|x)$ 在区间 $(0,\theta_0)$ 上单调增, 而在区间 $(\theta_0,1)$ 上单调减;

(d) 对于某个 θ_0, 函数 $\pi(\theta|x)$ 在区间 $(0,\theta_0)$ 上单调减, 而在区间 $(\theta_0,1)$ 上单调增.

相应地, HPD 区域 $J(x)$ 具有如下形式之一:

(a) $\theta<v(x)$;

(b) $\theta>v(x)$;

(c) $v_1(x)<\theta<v_2(x)$;

(d) $\theta<v_1(x)$ 或者 $\theta>v_2(x)$;

对于情形 (a) 和 (b), $v(x)$ 由区域的后验概率是 $1-\alpha$ 确定; 对于情形 (c) 和 (d), 为确定 $v_1(x)$ 和 $v_2(x)$, 则除了后验概率这一要求, 还要添加

$$\pi\big(v_1(x)|x\big)=\pi\big(v_2(x)|x\big)$$

的条件. 以情形 (c) 为例, 这两个条件是:

$$\frac{\Gamma(n+a+b)}{\Gamma(a+x)\Gamma(b+n-x)}\int_{v_1(x)}^{v_2(x)}t^{a+x-1}(1-t)^{b+n-x-1}dt=1-\alpha,$$

$$v_1(x)^{a+x-1}(1-v_1(x))^{b+n-x-1}=v_2(x)^{a+x-1}(1-v_2(x))^{b+n-x-1}.$$

例 4.12 设样本 $X_1,\cdots,X_n$ i.i.d., 服从正态分布 $N(\theta,\sigma^2)$. 当缺少关于参数的信息时, (θ,σ) 的先验分布 ξ 可取为

$$\xi(d\theta,d\sigma)=d\theta d\sigma/\sigma,\quad -\infty<\theta<\infty,\ \sigma>0,$$

要寻求在此先验分布之下 θ 和 σ 各自的 Bayes 可信区间.

给定 $x=(x_1,\cdots,x_n)$, (θ,σ) 的联合后验分布为

$$\pi(\theta,\sigma|x)d\theta d\sigma=A(x)\frac{1}{\sigma^{n+1}}\exp\Big\{-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(\theta-x_i)^2\Big\}d\theta d\sigma.$$

经过简单计算, θ 和 σ 的边缘后验密度分别具有如下形式:

$$\pi_1(\theta|x) = A_1(x)\Big[1+\frac{n(\theta-\bar{x})^2}{(n-1)s^2}\Big]^{-n/2}, \tag{4.20}$$

$$\pi_2(\sigma|x) = A_2(x)\sigma^{-n}\exp\Big\{-\frac{1}{2\sigma^2}(n-1)s^2\Big\}, \tag{4.21}$$

其中 $\bar{x}=\sum_{i=1}^n x_i/n, s^2=(n-1)^{-1}\sum_{i=1}^n(x_i-\bar{x})^2$.

对给定的 x, 作变换

$$\theta \mapsto T = \frac{\sqrt{n}(\theta-\bar{x})}{s},$$

由 θ 的后验密度 (4.20), 即得给定 x 时 T 的条件密度

$$g(t\ |x) = B_n\Big(1+\frac{t^2}{n-1}\Big)^{-n/2}, \quad B_n = \frac{\Gamma(n/2)}{\sqrt{(n-1)\pi}\Gamma((n-1)/2)},$$

即对任何给定的 x, $T\ |_x \sim t_{n-1}$.

给定 x, 为得到满足 (4.19) 的 HPD 区域 $\pi_1(\theta|x)\geqslant c(x)$, 只需选取 $a(x)$ 使等价的区域 $|T|\leqslant a(x)$ 满足

$$P\{|T|\leqslant a(x)\ |x\} = P\{\pi_1(\theta\ |x)\geqslant c(x)\ |x\} = 1-\alpha,$$

由 $T\ |_x \sim t_{n-1}$, 可取 $a(x)=t_{n-1}(\alpha/2)$ 与 x 无关, 故所要的 HPD 区间和例 4.1 中所得置信区间 $\bar{X}\pm t_{n-1}(\alpha/2)S/\sqrt{n}$ 形式上一致, 但它们是在不同的假设下导出的, 统计解释也不一样.

为构建 σ 的 Bayes 可信区间, 对给定的 x, 作变换

$$\sigma \mapsto V = \sigma/s,$$

由 (4.21), 对任何给定的 x, V 的条件密度为

$$h(v\ |x) = C_n v^{-n}\exp\{-(n-1)/(2v^2)\}I(v>0) := h(v), \tag{4.22}$$

其中 $C_n = 2[(n-1)/2]^{(n-1)/2}/\Gamma\big((n-1)/2\big)$. 因 $\pi_2(\sigma\ |x)$ 具有例 4.11 中形式 (c), 故 σ 的可信系数为 $1-\alpha$ 的 HPD 区域形如 $c_1(x)\leqslant\sigma\leqslant c_2(x)$, 这个 HPD 区域可等价地写成 $as\leqslant\sigma\leqslant bs$ 即 $a\leqslant V\leqslant b$ 的形式, 其中 a 和 b 满足以下两个条件:

$$\int_a^b h(v)dv \equiv \int_{(n-1)/b^2}^{(n-1)/a^2} k_{n-1}(u)du = 1-\alpha,$$

$$h(a)=h(b), \text{即 } a^{-n}\exp\{-(n-1)/(2a^2)\} = b^{-n}\exp\{-(n-1)/(2b^2)\}, \tag{4.23}$$

其中 k_{n-1} 为分布 χ^2_{n-1} 的密度函数. 用定理 4.6 中的方法, 在 σ 的形如 $[c_1S,\ c_2S]$ 的 $1-\alpha$ 置信区间中, 不难求得最短的那个置信区间, σ 的这个可信系数为 $1-\alpha$ 的 HPD 区间在形式上与它完全一样.

4.5 信仰推断法

信仰推断法 (fiducial inference) 是 Fisher 在 20 世纪 30 年代初期提出的一种统计方法. 我们可以说, 这一方法的要旨在于, 在不涉及参数的先验分布的情况下, 根

据样本 x 定出参数的一个分布. 这个分布在数学上具有概率分布的性质, 但不能给予频率解释. 它并不是因为参数具有随机变量的特征带来的, 而只是表达了参数落在种种范围内的 “可信任程度”, 因而被称为 “信仰分布” (fiducial distribution). 应当指出, 这种 “可信任程度” 归根结底还是基于样本信息. 我们先通过一个例子来说明这种方法的基本思想.

例 4.13 设 $X_1,\cdots,X_n$ 为从正态总体 $N(\theta,1)$ 中抽取的 i.i.d. 样本, 要对未知参数 θ 作出推断.

考虑充分统计量 $\bar{X}$, 其分布为

$$\bar{X}-\theta\sim N(0,1/n). \tag{4.24}$$

撇开 Bayes 统计不谈, 按照我们一贯的看法, θ 是常数, 无随机性, 随机变量是 $\bar{X}$. Fisher 把上式反过来看, 把 $\bar{X}$ 看成固定的常数, 而认为 (4.24) 确定了 θ 的一个分布, 即 $N(\bar{X},1/n)$, 称之为 θ 的**信仰分布**. 直观上可作如下解释: 在抽样之前我们对 θ 一无所知, 在抽得样本 $X_1,\cdots,X_n$ 之后, 对 θ 的情况有了一些信息, 因而对 θ 取不同值的 “信仰程度” 有所不同, 例如, 对于 θ 落在各等长小区间中的信仰程度, 随区间中点与 $\bar{X}$ 距离的增大而呈下降的趋势. 分布 $N(\bar{X},1/n)$ 对这种信仰程度给出了量的刻画. 据此信仰分布即可构建 θ 的信仰区间或信仰上、下限, 其**信仰水平或信仰系数**均可仿照置信区间或 Bayes 可信区间的情形定义. 这样我们即可得到 θ 的一个**信仰系数为 $1-\alpha$ 的区间估计** $[\bar{X}-u_{\alpha/2}/\sqrt{n},\ \bar{X}+u_{\alpha/2}/\sqrt{n}]$. 虽然在形式上这个信仰区间与常用的置信系数为 $1-\alpha$ 置信区间完全一样, 但它们的统计解释却大不相同.

例 4.14 设样本 $X_1,\cdots,X_n$ i.i.d., $X_1\sim U(0,\theta),\theta>0$. 为了用信仰推断法作出 θ 的区间估计, 与例 4.13 类似, 我们从 θ 的充分统计量 $X_{(n)}=\max_{1\leqslant i\leqslant n}X_i$ 出发. 对 $\theta>0$, $X_{(n)}/\theta$ 的密度函数为 $p(w)=nw^{n-1}I(0<w<1)$, 与 θ 无关. 为了诱导出 θ 的信仰密度, 我们将 $X_{(n)}=x_{(n)}$ 视为常数, 而将 θ 看成随机变量, 并假定 $W:=x_{(n)}/\theta$ 仍具有密度 $p(w)$. 此时 θ 的密度函数即为

$$q(\theta)=p\Big(\frac{x_{(n)}}{\theta}\Big)\cdot\frac{x_{(n)}}{\theta^2}=\frac{nx_{(n)}^n}{\theta^{n+1}}I\big(x_{(n)}<\theta<\infty\big),$$

我们就把这一密度作为参数 θ 的信仰密度. 据此即可构建需要的信仰区间. 在此, 如果存在一个与样本 x 有关的区间, 信仰密度在区间内每点的取值都 $\geqslant$ 区间外的取值, 且具有指定的信仰系数, 则称之为**最高信仰密度区间** (highest fiducial density interval), 简记为 **HFD 区间**. 由于在本例中 $q(\theta)$ 关于 θ 严格减, 故据此信仰分布所得信仰系数为 $1-\alpha$ 的最短的信仰区间, 即为 HFD 区间 $[x_{(n)},b(x_{(n)})]$, 其中 $b=b(x_{(n)})$ 由条件

$$P(x_{(n)}\leqslant\theta\leqslant b)=\int_{x_{(n)}}^{b}\frac{nx_{(n)}^n}{\theta^{n+1}}=1-\alpha$$

确定, 这个 HFD 区间即为 $[X_{(n)},X_{(n)}/\alpha^{1/n}]$, 它在形式上与例 4.9 最后一段中得出的 θ 的 $1-\alpha$ UMA 置信区间一样.

在这两个例子中, 为了用信仰推断法构建区间估计, 我们都是利用模型的充分统计量构建了一个枢轴变量, 其分布与参数无关, 并以此分布诱导出参数的一个信仰分布. 下面我们用这种方法来构建正态总体的均值和标准差的区间估计.

例 4.15 设 $X_1,\cdots,X_n$ i.i.d., $X_1\sim N(\mu,\sigma^2)$, $\mu\in\mathbf{R}$ 和 $\sigma^2>0$ 都未知. $(\bar X,S^2)$ 是 (μ,σ^2) 的完全充分统计量, 且 $\bar X$ 和 S^2 独立, 其中

$$\bar X=\sum_{i=1}^n X_i/n,\quad S^2=\frac{1}{n-1}\sum_{i=1}^n(X_i-\bar X)^2.$$

为构建 μ 和 σ 的信仰区间, 记

$$\xi=\sqrt{n}(\bar X-\theta)/\sigma,\quad \eta=S/\sigma, \tag{4.25}$$

则 ξ 和 η 独立, 其联合密度为 $f(\xi,\,\eta)=f_1(\xi)f_2(\eta)$, 其中

$$f_1(\xi)=(2\pi)^{-1/2}\exp\{-\xi^2/2\},$$

$$f_2(\eta)=C_n\eta^{n-2}\exp\{-(n-1)\eta^2/2\}I(\eta>0),$$

而 $C_n=2\big((n-1)/2\big)^{(n-1)/2}/\Gamma\big((n-1)/2\big)$. (ξ,η) 即可作为枢轴变量, 其分布与参数 (θ,σ^2) 无关. 给定样本 $X=x$, 即给定 $(\bar X,S^2)=(\bar x,s^2)$, 我们把变量 (ξ,η) 看成变量 (θ,σ) 的函数, 仍赋予上式所示的联合密度 $f(\xi,\eta)=f_1(\xi)f_2(\eta)$, 由此即可诱导出 (θ,σ) 的联合信仰密度

$$q(\theta,\sigma)=\sqrt{n}C_n\frac{s^{n-1}}{\sqrt{2\pi}\sigma^{n+1}}\exp\Big\{-\frac{n(\bar x-\theta)^2+(n-1)s^2}{2\sigma^2}\Big\}I(\sigma>0) \tag{4.26}$$

和 σ 的边缘信仰密度

$$q_2(\sigma)=C_n s^{n-1}\sigma^{-n}\exp\{-(n-1)s^2/(2\sigma^2)\}I(\sigma>0). \tag{4.27}$$

把 (4.26) 对 σ 积分, 可得 θ 的边缘信仰密度, 结果表明

$$\sqrt{n}(\theta-\bar x)/s\sim t_{n-1}. \tag{4.28}$$

由此可以作出 θ 的信仰区间 $\bar X\pm t_{n-1}(\alpha/2)S/\sqrt{n}$, 信仰系数为 $1-\alpha$. 这种区间估计形式上与 Neyman 意义的 t-区间估计完全一样, 但解释不同: 理论上后者允许有频率解释, 而前者则表明在有了样本 $X_1,\cdots,X_n$ 之后, 我们对 "上述区间包含 θ" 这件事的信仰程度为 $1-\alpha$. 但就本例而言, 若从实用的角度, 则常被视为用不同的方法得出了同样的结果.

由 (4.27), $V:=\sigma/s$ 的信仰密度为

$$h(v)=C_n v^{-n}\exp\{-(n-1)/(2v^2)\}I(v>0), \tag{4.29}$$

形式上与 (4.22) 一样. 由此即得参数 σ 的信仰系数为 $1-\alpha$ 的 HFD 区间为 $[aS,bS]$, 其中 $0<a<b$ 由方程 (4.23) 确定, 因而在本例中得到的信仰系数为 $1-\alpha$ 的 θ 和 σ 的两个 HFD 区间, 形式上与在例 4.12 中的无信息先验分布之下得到的可信系数为 $1-\alpha$ 的 HPD 区间完全一样. 正如例 4.12 指出的, σ 这个 HFD 区间还与 σ 的形如 $[c_1S,c_2S]$ 的 $1-\alpha$ 置信区间中最短的那个完全一样.

在上述例子中, Fisher 的信仰推断法给出的参数的区间估计, 与传统的置信区间是一致的. 这种形式上的一致性表明, 虽然它们是在不同的意义下得到的, 其结果

的解释也不一样, 但在这些模型中并没有提供新的估计. 然而将其用于处理著名的 Behrens-Fisher 问题, 却有其新鲜之处. 鉴于在这个异方差的两样本正态模型中, 用传统方法寻求均值差的确切的置信区间的努力, 一直未能给出令人满意的处理方法, 说这个新方法在当时引起了轰动效应也不为过.

例 4.16 设样本 $X_1,\cdots,X_m$ i.i.d., $Y_1,\cdots,Y_n$ i.i.d., 且全体独立, $X_1\sim N(\theta_1,\sigma_1^2)$, $Y_1\sim N(\theta_2,\sigma_2^2)$, 其中 $-\infty<\theta_1,\theta_2<\infty,\sigma_1^2,\sigma_2^2>0$, $m,n\geqslant 2$. 我们要用信仰推断法构建 $\theta_2-\theta_1$ 的区间估计. 为此, 记

$$S_1=\frac{1}{\sqrt{m}}\Big(\frac{1}{m-1}\sum_{i=1}^{m}(X_i-\bar{X})^2\Big)^{1/2},$$
$$S_2=\frac{1}{\sqrt{n}}\Big(\frac{1}{n-1}\sum_{i=1}^{n}(Y_i-\bar{Y})^2\Big)^{1/2}.$$

将 $\bar{X},\bar{Y},S_1,S_2$ 的取值分别记为 $\bar{x},\bar{y},s_1,s_2$ 并视为常数, 则在例 4.15 中业已证明,

$$\theta_1\overset{d.}{=}\bar{x}-s_1T_1,\quad \theta_2\overset{d.}{=}\bar{y}-s_2T_2,$$

此处, T_1,T_2 独立, $T_1\sim t_{m-1},T_2\sim t_{n-1}$, "$\overset{d.}{=}$" 表示等号两边有相同的信仰分布. 令 $\psi=\arccos(s_2/\sqrt{s_1^2+s_2^2})$, 则

$$(\theta_2-\theta_1)-(\bar{y}-\bar{x})\overset{d.}{=}s_2T_2-s_1T_1=(s_1^2+s_2^2)^{1/2}W,\quad W=T_2\cos\psi-T_1\sin\psi,$$

其中, W 的密度函数关于零点对称, 分布涉及三个参数 ψ,m 和 n. 设 $\zeta_{\alpha/2}$ 为该分布的上 $\alpha/2$ 分位数, 则可取 $\bar{Y}-\bar{X}\pm\zeta_{\alpha/2}(S_1^2+S_2^2)^{1/2}$ 作为 $\theta_2-\theta_1$ 的区间估计, 其信仰系数为 $1-\alpha$. 在 Fisher 和 Yates (1963) 中载有该分布的表.

这个方法的成功之处, 在于给出的区间确切地具有指定的信仰系数. 而 Neyman 方法无法给出指定置信系数的置信区间. 对于这种解法, Neyman 的评论主要有两条: 一是信仰分布的导出, 乃至其概念本身, 都缺乏根据; 二是他通过计算证明, 这样得出的区间估计并不具有在他的意义下的置信系数 $1-\alpha$. 他选取了 $m=12,n=6$ 和 $\alpha=0.05$, 算出上述信仰区间包含 $\theta_2-\theta_1$ 的概率与比值 $\rho=\sigma_1/\sigma_2$ 有关, 当 $\rho=0.1,1$ 和 10 时, 这个概率分别为 0.966, 0.960 和 0.934. 基于概率的频率解释, 提出第一条评论甚属自然. 至于计算结果, 信仰推断论者认为, 其意义只是表明信仰系数与置信系数的确不是一回事, 而一种体系只要自身相容且在实用中有良好表现, 就能自成学派, 而不取决于它是否与现存理论相一致. 事实上, 令人感兴趣的是, 即使以 Neyman 的置信系数为参照, 在 m,n 不大且有较大差距, 而 ρ 又有较大跨度的情况下, 上述信仰区间的信仰系数与置信系数依然非常接近, 说明 Fisher 方法还是令人放心的.

参照近期一些文献的研究, 我们将上述方法的一般提法归纳如下. 假定样本空间 $(\mathcal{X},\mathcal{B})$ 是欧氏的, 参数空间 Θ 是欧氏空间 $\mathbf{R}^p$ 的一个开子集. 基于样本 X, 我们可以构建一个 p-维枢轴变量 $h(X,\theta)$, 其密度函数与 $\theta\in\Theta$ 无关, 记为 $p(\cdot)$. 为了导出 θ 的信仰密度, 我们视 $X=x$ 为常量, 并按照有关文献中的做法, 引进一个 p-维随机变量 Z, 并假定 $W:=h(x,Z)$ 仍有密度函数 $p(\cdot)$. 在这一机制下, 若对给定的 x, $h:Z\mapsto W$ 构成 Z 和 W 的值域空间上的连续可微的一一映照, 其 Jacobi 矩阵非奇

异, 则 Z 的密度函数为

$$f(z;x)=p\big(h(x,z)\big)\Big|\det\Big(\frac{\partial h(x,z)}{\partial z^{\tau}}\Big)\Big|. \tag{4.30}$$

Z 的这个分布即可视为 θ 的一个信仰分布. 有了这个信仰分布, 即可构建 θ 的信仰系数为 $1-\alpha$ 的信仰区域. 但这样的信仰区域有多个, 通常推荐的是满足条件

$$J(x):=\{z: f(z;x)\geqslant c(x)\}\ \text{ 且 }\ \int_{J(x)} f(z;x)dz=1-\alpha, x\in\mathcal{X} \tag{4.31}$$

的**最高信仰密度区域** (highest fiducial density region), 简称 **HFD 区域** $J(X)$. 为了方便, 我们常常是先找参数 θ 的一个点估计 $T=T(X)$, 再构建枢轴变量. 选用的 T 不同, 所得信仰密度也会不一样.

为了使得 HFD 区域的构建更加方便, 下面介绍文献中引进的多维密度函数的一种垂直表示方法. 为此, 设 Z 为一 p 维随机变量, 具有密度函数 f. 记

$$D_f=\{(z^{\tau},z_{p+1})^{\tau}: 0\leqslant z_{p+1}\leqslant f(z), z\in\mathbf{R}^p, z_{p+1}\in\mathbf{R}\},$$
$$D_f(v)=\{z\in\mathbf{R}^p: (z^{\tau},v)^{\tau}\in D_f\},\ 0<v\leqslant f_0=\sup_z f(z),$$

分别称它们为函数 f 的图下区域及其在 $z_{p+1}=v$ 处的截口. Troutt (1991) 证明了如下的定理.

定理 4.7 *若 $V:=f(Z)$ 有密度函数 g, 则*

$$g(v)=-v\frac{dL_p\big(D_f(v)\big)}{dv}, \tag{4.32}$$

其中 L_p 为 p-维 Lebesgue 测度.

为了不打断整体思路, 这一定理的证明将放到本节最后进行.

现在回到 (4.31). 记住在 (4.31) 中, x 为给定的常数. 由假定, Z 的密度函数为 $f(z;x)$, 按定理 4.7, 可算得 $V_x:=f(Z;x)$ 的密度函数 $g(v;x)$, 而 HFD 区域 $J(X)$ 即可简单地由方程

$$J(x):=\{z: f(z;x)\geqslant c(x)\}\ \text{ 且 }\ \int_{c(x)}^{f_{0,x}} g(v;x)dv=1-\alpha, x\in\mathcal{X} \tag{4.33}$$

确定, 此处 $f_{0,x}:=\sup_z f(z;x)$.

Fang, Yang 和 Kotz (2001) 将 Troutt 在定理 4.7 中的结果称为 **I 型垂直密度表示** (vertical density representation), 简称 **I 型 VDR**, 并提出了基于几何概率的 **II 型 VDR**, 这就是下面的定理.

定理 4.8 *以下三个命题等价:*

(a) *随机向量 $(Z^{\tau},Z_{p+1})^{\tau}$ 服从 D_f 上的均匀分布;*

(b) *Z_{p+1} 的边缘密度为*

$$f_{p+1}(v)=\begin{cases} L_p\big(D_f(v)\big), & \text{若 } 0<v<f_0=\sup\limits_z f(z),\\ 0, & \text{否则,}\end{cases}$$

且在给定 $Z_{p+1}=v$ 的条件下, Z 的条件密度为 $D_f(v)$ 上的均匀分布;

(c) Z 的边缘密度为 f, 且在给定 $Z=z$ 的条件下, Z_{p+1} 的条件分布为区间 $\big(0,f(z)\big)$ 上的均匀分布.

注意到 $L_{p+1}\big(D_f\big)=\int_{\mathbf{R}^p} f(z)dz=1$, 由联合密度、边缘密度、条件密度之间的关系, 读者不难验证这一定理成立.

下面我们在定理 4.8 的基础上, 给出定理 4.7 的证明.

证明 设 $0<v<f_0$, 则由定理 4.8,

$$\begin{aligned}P\{f(Z)\leqslant v\}&=P\{f(Z)\leqslant v,(Z^\tau,Z_{p+1})^\tau\in D_f\}\\&=L_{p+1}\{(z^\tau,z_{p+1})^\tau\in D_f,f(z)\leqslant v\}\\&=L_{p+1}\{[(z^\tau,z_{p+1})^\tau\in D_f,z_{p+1}\leqslant v]-D_f(v)\times(0,v]\}\\&=L_{p+1}\{(z^\tau,z_{p+1})^\tau\in D_f,z_{p+1}\leqslant v\}-vL_p\big(D_f(v)\big)\\&=P\{Z_{p+1}\leqslant v\}-vL_p\big(D_f(v)\big).\end{aligned}$$

由 $f(Z)$ 有密度 g, Z_{p+1} 有密度 f_{p+1}, 知上式两边可对 v 求导, 故

$$g(v)=L_p\big(D_f(v)\big)-\frac{d\big[vL_p\big(D_f(v)\big)\big]}{dv}=-v\frac{dL_p\big(D_f(v)\big)}{dv},\quad \text{a.e. } L,$$

此处 L 为 $\mathbf{R}$ 上的 Lebesgue 测度. ■

由此还可知道: $f(Z)$ **有密度, 当且仅当** $L_p\big(D_f(v)\big)$ **绝对连续.**

为了完善和发展 Fisher 的信仰推断法, 先后有不少学者作出了他们的贡献. 由于本书的篇幅有限, 我们只能限于上面这样一些介绍. 关于这一理论的近期工作, 涉及参数的随机估计、置信分布等概念, 以及将垂直密度表示用于构建多维参数的信仰区域的 VDR 方法和 VDR 置信域的丰富内容, 读者可参看 Xie 和 Singh (2013), 杨振海 (2014) (专著) 以及其中有关的参考文献. 关于这一方法的哲学层面的评论, 可参看陈希孺 (1999, 2009). 本书作者参考了这些文献, 从中受益匪浅.

4.6 习题和补充

在以下的陈述中, $I(A)$ 表示集合 A 的示性函数, $0<1-\alpha<1$ 是事先指定的一个正数, "θ 的置信水平为 $1-\alpha$ 的置信区间" 简称为 "θ 的 $1-\alpha$ 置信区间".

$$X_{(1)}:=\min_{1\leqslant i\leqslant n}X_i,\qquad X_{(n)}:=\max_{1\leqslant i\leqslant n}X_i,$$

对于小写字母也常用类似的记号.

4.1 在例 4.4 中, 取 $\theta/X_{(1)}$ 作为枢轴变量, 利用定理 4.6, 在 θ 的形如 $[c_1X_{(1)},c_2X_{(1)}]$ 的 $1-\alpha$ 置信区间中, 寻求长度最小的置信区间.

4.2 在例 4.2 中, 取 $\theta/X_{(n)}$ 作为枢轴变量, 利用定理 4.6, 在 θ 的形如 $[c_1X_{(n)},c_2X_{(n)}]$ 的 $1-\alpha$ 置信区间中, 寻求长度最小的置信区间.

4.3 利用定理 4.6, 重新研究例 4.1 中有关的最短置信区间问题:

(1) 在 μ 的形如 $[\bar{X}+c_1S, \bar{X}+c_2S]$ 的 $1-\alpha$ 置信区间中, 寻求长度最小的置信区间;

(2) 在 σ^2 的形如 $[c_1S^2, c_2S^2]$ 的 $1-\alpha$ 置信区间中, 寻求长度最小的置信区间.

4.4 设 $X_1, \cdots, X_n$ i.i.d., $X_1 \sim E(0,\theta), \theta>0$, $X_{(1)} \leqslant X_{(2)} \leqslant \cdots \leqslant X_{(n)}$ 为它的次序统计量. 假定我们只能观测到 $(X_{(1)}, X_{(2)}, \cdots, X_{(r)})$, 其中 $r \geqslant 2$ 是事先指定的一个整数. 令 $T=\sum_{i=1}^{r} X_{(i)}+(n-r)X_{(r)}$, 则 T 为完全充分统计量, $2T/\theta \sim \chi^2_{2r}$. 试在 θ 的形如 $[aT,\ bT]$ 的 $1-\alpha$ 置信区间中, 寻求长度最小的置信区间.

4.5 设 $X_1, \cdots, X_n$ i.i.d., $X_1 \sim f(x-\mu)dx$, 其中 f 已知, μ 未知. 考虑 μ 的形如 $[\bar{X}-c_1, \bar{X}+c_1]$ 和 $[X_1-c_2, X_1+c_2]$ 的置信系数为 $1-\alpha$ 的置信区间.

(1) 设 f 是 $N(0,1)$ 的密度函数, 证明前者的长度较后者为小.

(2) 设 f 是标准 Cauchy 分布的密度函数, 证明两个置信区间具有相同的长度.

4.6 设 $X_1, \cdots, X_n$ 是来自指数分布 $E(\theta,\theta)$ 的 i.i.d. 样本, 其中 $\theta>0$ 未知, $n \geqslant 2$.

(1) 分别以 $a_{n,\alpha}$ 和 $b_{n,\alpha}$ 记枢轴变量 $X_{(1)}/\theta-1$ 和 $\bar{X}/\theta-1$ 的分布的 α 分位数, 由不等式 $a_{n,\alpha/2} \leqslant X_{(1)}/\theta-1 \leqslant a_{n,1-\alpha/2}$ 和 $b_{n,\alpha/2} \leqslant \bar{X}/\theta-1 \leqslant b_{n,1-\alpha/2}$, 可以得到 θ 的两个不同的 $1-\alpha$ 置信区间, 给出这两个区间的明确形式.

(2) 比较这两个置信区间的长度.

4.7 设 $X_1, \cdots, X_n$ 是来自指数分布 $E(\theta,\theta)$ 的 i.i.d. 样本, 其中 $\theta>0$ 未知. 以 $\theta/X_{(1)}$ 为枢轴变量, 分别对 $n=1$ 和 $n \geqslant 2$ 两种情形, 在 θ 的形如 $[aX_{(1)}, bX_{(1)}]$ 的 $1-\alpha$ 置信区间中, 寻求长度最小的置信区间.

*4.8 对 Behrens-Fisher 问题, 记 $S_1^2=\sum_{i=1}^{m}(X_i-\bar{X})^2$, $S_2^2=\sum_{i=1}^{n}(Y_i-\bar{Y})^2$. 若能找到函数 $f(s_1,s_2)$, 使

$$\frac{(\bar{Y}-\bar{X})-(\theta_2-\theta_1)}{f(S_1,S_2)}$$

的分布与 σ_1, σ_2 无关, 则可据以建立 $\theta_2-\theta_1$ 的一个相似区间. 证明: 若将 f 限制在 $\{f : \mathrm{E}_{\sigma_1,\sigma_2}[f^2(S_1,S_2)]<\infty$, 对所有 $\sigma_1>0,\ \sigma_2>0\}$ 的函数类中, 则这样的 f 不存在.

4.9 设 $X_1, \cdots, X_n$ 为来自 Possion 分布总体 $P(\lambda)$ 的 i.i.d. 样本, 其中 $\lambda>0$ 未知.

(1) 构建 λ 的渐近置信系数为 $1-\alpha$ 的大样本置信限和置信区间.

(2) 参考例 4.6, 构建 λ 的 $1-\alpha$ 置信限和置信区间.

4.10 设 $X_1, \cdots, X_n$ 是来自指数总体 $\{E(\theta,1), \theta \in \mathbf{R}\}$ 的 i.i.d. 样本, 其共同的分布密度为

$$f(x;\theta)=\begin{cases}\exp\{-(x-\theta)\}, & x \geqslant \theta,\\ 0, & x<\theta.\end{cases}$$

试构建 θ 的基于 $X_{(1)}$ 的 $1-\alpha$ 置信区间.

4.11 设总体 X 服从点集 $\{1,2,\cdots,N\}$ 上的均匀分布, 其中 N 为未知参数, $X_1, \cdots, X_n$

是来自该分布的 i.i.d. 样本. 试基于 $X_{(n)}$ 构建 N 的 $1-\alpha$ 置信上限.

4.12 设 $X_1,\cdots,X_m$ 和 $Y_1,\cdots,Y_n$ 分别为来自 $N(\mu,\sigma_1^2)$ 和 $N(c\mu,\sigma_2^2)$ 的 i.i.d. 样本, 其中 c, σ_1^2 和 σ_2^2 已知, μ 未知, $c\neq 0$. 又设两样本相互独立.

(1) 找出 μ 的 UMVU 估计.

(2) 构建 μ 基于这一点估计的 $1-\alpha$ 置信区间.

4.13 设 $X_1,\cdots,X_n$ 是来自均匀分布 $U(\theta-a,\theta+a)$ 的 i.i.d. 样本, $-\infty<\theta<\infty, a>0$.

(1) 构建 a 的基于充分统计量 $(X_{(1)},X_{(n)})$ 的 $1-\alpha$ 置信区间;

(2) 求 $[X_{(1)}+X_{(n)}-(2\theta)]/(X_{(n)}-X_{(1)})$ 的分布, 并构建 θ 的 $1-\alpha$ 置信区间.

4.14 设 $X_1,\cdots,X_m$ 和 $Y_1,\cdots,Y_n$ 分别为来自均匀分布 $U(0,\theta_1)$ 和 $U(0,\theta_2)$ 的 i.i.d. 样本, 其中 $\theta_1>0,\theta_2>0$ 皆未知, 且两样本独立. 试构建 θ_2/θ_1 的基于 $(X_{(m)},Y_{(n)})$ 的 $1-\alpha$ 置信区间.

4.15 设样本 $X\sim P_\theta$, 要构建 θ 的一个置信区间. 假定 $T(X)$ 是一个实值统计量, 其概率密度函数为 $f_\theta(t)$, 且对每个给定的 θ, $R(t,\theta)$ 是 t 的单调函数. 证明如果存在一个函数 g, 使

$$f_\theta(t)=g(R(t,\theta))\left|\det\left[\frac{\partial R(t,\theta)}{\partial t}\right]\right|,$$

则 $R(T(X),\theta)$ 可以取为一个枢轴变量.

4.16 设 $X_1,\cdots,X_n$ 是来自 $N(\theta,\theta)$ 的 i.i.d. 样本, 其中 $\theta>0$ 未知. 试构建一个枢轴变量, 进而构建 θ 的一个 $1-\alpha$ 置信区间.

4.17 (**多个参数的同时区间估计**) 设样本 $X_{i1},\cdots,X_{in}$ 为取自正态 $N(\theta_i,\sigma^2)$ 的 i.i.d. 样本, θ_i,σ 均为未知, $i=1,\cdots,k$, 且 k 个样本相互独立. 要同时构建 $\theta_1,\cdots,\theta_k$ 的区间估计.y

为此, 记 $\bar X_i=\sum_{j=1}^n X_{ij}/n, S_i^2=\sum_{j=1}^n(X_{ij}-\bar X_i)^2, i=1,\cdots,k$, $S^2=\sum_{i=1}^k S_i^2$, 则 $\bar X_1,\cdots,\bar X_k,S^2$ 相互独立. 记 $m=k(n-1)$. 如能确定常数 c 使事件

$$\left\{\frac{\max_{1\leqslant i\leqslant k}|\sqrt{n}(\bar X_i-\theta_i)|}{S/\sqrt{m}}\leqslant c\right\}$$

在参数 $\theta:=(\theta_1,\cdots,\theta_k,\sigma)$ 下都有指定的概率 $1-\alpha$, 则我们即得形如

$$\bar X_i-cS/\sqrt{nm}\leqslant\theta_i\leqslant\bar X_i+cS/\sqrt{nm},\quad i=1,\cdots,k$$

的相似的同时区间估计, 其置信水平为 $1-\alpha$. 所要的 c 即为变量

$$\xi=\frac{\max_{1\leqslant i\leqslant k}|Z_i|}{\sqrt{W_m/m}}$$

的上 α 分位数, 此处 $Z_1,\cdots,Z_k,W_m$ 相互独立, $Z_i\sim N(0,1), W_m\sim\chi_m^2$. 对 α,n,m 的一些值, Hahn 和 Hendrickson (1971) 制作的用表给出了此分布的上 α 分位数.

[此题和下一题也可参看 陈希孺 (1999, 2009) 例 7.5.]

4.18 (**多个参数的同时区间估计**) 续接上题. 早在 1960 年, Harter 就研究了如何对所有的 $\theta_i-\theta_j$ 构建形如

$$\bar{X}_i-\bar{X}_j-cS/\sqrt{nm}\leqslant\theta_i-\theta_j\leqslant\bar{X}_i-\bar{X}_j-cS/\sqrt{nm},\quad 1\leqslant i\leqslant j\leqslant k$$

的同时区间估计的问题, 并制作了有关的用表. 参看 Harter (1960). 试写出如何得出上述联合区间估计的机理.

4.19 设 $X_1,\cdots,X_m$ 和 $Y_1,\cdots,Y_n$ 分别是从来自 $N(a,\sigma^2)$ 和 $N(a,\rho\sigma^2)$ 的 i.i.d. 样本, 两样本整体独立, $\rho>0$ 已知, a 和 σ 未知. 求 a 的 $1-\alpha$ UMAU 置信区间.
[参看第三章习题 3.22.]

4.20 设 $X_1,\cdots,X_n$ 是来自 Pareto 分布总体的 i.i.d. 样本. Pareto 分布的密度函数为

$$f(x;\ \theta,\beta)=\beta\theta^{\beta}x^{-(\beta+1)}I(x>\theta),\ \theta>0,\beta>0,$$

其中 $\beta=2$ 已知. 试构建 θ 的 $1-\alpha$ UMA 置信上限.

4.21 设 $X_1,\cdots,X_n$ 是来自指数分布 $E(\mu,\theta)$ 的 i.i.d. 样本, $\mu\in\mathbf{R},\theta>0$. X_1 的分布为

$$f(x;\ \mu,\theta)dx=\frac{1}{\theta}\exp\Big\{-\frac{x-\mu}{\theta}\Big\}I(x>\mu)dx.$$

(1) 试构建 μ 的 $1-\alpha$ UMAU 置信下限;
(2) 试构建 θ 的 $1-\alpha$ UMAU 置信区间.

4.22 设 $X_1,\cdots,X_n$ 是来自分布 $f(x,\theta)dx=\theta x^{\theta-1}I(0<x<1)dx$ 的 i.i.d. 样本, 其中 $\theta>0$ 未知.
(1) 构建 θ 的基于充分统计量的 $1-\alpha$ 置信区间.
(2) 求 θ 的 $1-\alpha$ UMAU 置信区间.

*4.23 设 X 为取自分布族 $\mathcal{P}:=\{P_\theta,-\infty<\theta<\infty\}$ 的单个观测样本, 其中

$$P_\theta=\begin{cases}N(\theta-1,1), & \theta<0,\\ N(0,1), & \theta=0,\\ N(\theta+1,1), & \theta>0.\end{cases}$$

(1) 证明分布族 $\mathcal{P}$ 是一个单调似然比分布族.
(2) 求 θ 的置信水平为 $1-\alpha$ 的置信下限.
[提示: 如在 (θ,x) 平面上标出有关假设检验的接受域即可一目了然.]

4.24 设 $X_1,\cdots,X_n$ 是 i.i.d. Bernoulli 随机变量, $P(X_i=1)=1-P(X_i=0)=p$. 取 Beta 分布作为 p 的先验分布. 试构建 p 的可信系数为 $1-\alpha$ 的 HPD 可信集.

4.25 重新考察例 4.2.
(1) 引进 Pareto 分布 $Pa(\zeta,\gamma)$ 作为 θ 的先验分布:

$$\pi(\theta)d\theta=\gamma\zeta^{\gamma}\theta^{-(\gamma+1)}I(\theta>\zeta)d\theta,\ \zeta>0,\gamma>0.$$

求 θ 的可信系数为 $1-\alpha$ 的 HPD 可信区间.
(2) 问对 θ 应取何种广义先验分布, 其可信系数为 $1-\alpha$ 的 HPD 可信区间与例 4.2 中所得置信区间形式上完全一致?

4.26 重新考察例 4.4.

(1) 引进下面的分布作为 θ 的先验分布:

$$\pi(\theta)d\theta = \gamma\theta^{\gamma-1}/\zeta^{\gamma} I(0<\theta<\zeta), \quad \gamma>0, \zeta>0.$$

求 θ 的可信系数为 $1-\alpha$ 的 HPD 可信区间.

(2) 问对 θ 应取何种广义先验分布, 其可信系数为 $1-\alpha$ 的 HPD 可信区间与例 4.4 中所得置信区间形式上完全一致?

4.27 设 $X_1, \cdots, X_n$ i.i.d., $X_1 \sim E(0,\theta), \theta>0$, $X_{(1)} \leqslant X_{(2)} \leqslant \cdots \leqslant X_{(n)}$ 为它的次序统计量, 而我们只能观测到 $(X_{(1)}, \cdots, X_{(r)})$, 其中 $r \geqslant 2$ 是事先指定的一个整数. 将 θ 的先验分布取为如下的逆 Gamma 分布 $\Gamma^{-1}(\kappa, 1/\beta)$:

$$\pi(\theta)d\theta = \frac{\beta^{\kappa}}{\Gamma(\kappa)}\theta^{-(\kappa+1)}\mathrm{e}^{-\beta/\theta} I(\theta>0), \ \kappa>0, \beta>0.$$

试求 θ 的可信系数为 $1-\alpha$ 的 HPD 可信区间.

4.28 设 $X_1, \cdots, X_n$ 为从密度函数为 $f(x,\lambda) = \lambda\mathrm{e}^{-\lambda x} I(x>0)$ 的指数分布中抽取的 i.i.d. 样本, 试导出 λ 的信仰分布, 进而给出信仰系数为 $1-\alpha$ 的最短区间估计. [提示: 利用充分统计量 $\bar{X}$.]

4.29 接习题 4.21, 试以相互独立的充分统计量 $(X_{(1)}, \bar{X}-X_{(1)})$ 为出发点, 导出 (μ,θ) 的信仰分布, 进而分别给出 μ 和 θ 的信仰系数为 $1-\alpha$ 的信仰区间. 说明这样得到的信仰区间是否与某个 $1-\alpha$ 置信区间一致.

第五章

统计渐近理论

本章主要介绍渐近统计即大样本统计的基本理论. 为方便计, 容量为 n 的样本 $X_1, \cdots, X_n$ 简记为 X^n. 大样本统计主要研究两个层面的问题. 其一, 设对一个统计问题已有一个对充分大的样本量 n 都适用的统计方法, 其固定样本情形的某种统计性质难以求得, 需要研究的是 $n \to \infty$ 时该方法的有关渐近性质; 其二, 设从某种准则或某种直观考虑出发, 认为对某个统计问题, 应采用某种统计量 $g(X^n)$, 要求研究这种统计量的某种极限性质, 并据此建立一种当样本量 n 充分大时近似适用的统计方法. 在这种意义下建立的统计理论和方法, 称为大样本理论和方法. 反之, 如果一种统计方法, 无论是其方法的构建还是统计性质的建立, 都不诉求于样本量 $n \to \infty$ 的极限过程, 就称之为小样本方法, 与之相应的统计理论就称为小样本理论.

本章的大部分篇幅用于讨论大样本估计, 只用少量篇幅介绍大样本检验. 大样本估计的内容主要有两个方面. 一是研究点估计序列的若干基本的渐近性质, 主要包括相合性和渐近正态性. 进一步, 我们将在更广的范围内研究一些重要统计量序列的渐近分布理论. 二是研究以大样本性质为依托的一些重要估计, 例如 **M-估计和极大似然估计** (Maximum Likelihood Estimate, 简记为 MLE). 关于内容的安排顺序, 在介绍 MLE 的经典极限理论和 M-估计的渐近理论之后, 我们将在 M-估计的框架下重访 MLE, 介绍 Le Cam 等人不同的研究视角和由此得出的精妙结果. 考虑到在理论和方法等方面的连贯性, 我们紧跟着讲述似然比检验等大样本检验的内容, 而将次序统计量安排在后面的 5.6 节. 最后一节则是对回归分析的一个科普性的简介.

另外, 本章涉及的随机元序列的各种收敛性的概念和记号, 在本书附录 A.2, A.3 和 A.4 中都有详细的介绍, 为避免重复, 就不再一一交代, 读者可参看附录中的有关部分.

5.1 估计的相合性和渐近正态性

5.1.1 基本概念

设总体的随机元 $X \sim P_\theta, \theta \in \Theta$, $g(\theta)$ 为定义于参数空间 Θ 上的实值函数, 参数 θ 未知而 g 已知, 当样本为 $X^n := (X_1, \cdots, X_n)$ 时用 $T_n = T(X^n)$ 去估计它. 对一般情形而言, 有关估计仅当 n 超过某个 n_0 时才存在. 为简单计, $n > n_0$ 一语常被略去. 下面我们对估计序列引进渐近无偏性、相合性, 并对一般的统计量序列引进渐近正态性的概念.

$g(\theta)$ 的一列估计 T_n 称为是**渐近无偏**的, 如果对每个 $\theta \in \Theta$, 都有 $\mathrm{E}_\theta(T_n) \to g(\theta)$, $n \to \infty$.

T_n 称为是**弱相合**的, 如果对每个 $\theta \in \Theta$, 有 $T_n \xrightarrow{P_\theta} g(\theta), n \to \infty$.

T_n 称为是**强相合**的, 如果对每个 $\theta \in \Theta$, 有 $P_\theta\{T_n \to g(\theta)\} = 1$.

T_n 称为是 r 阶**平均相合**的, 如果对每个 $\theta \in \Theta$, 有 $\mathrm{E}_\theta|T_n - g(\theta)|^r \to 0$, 其中 r 为一正数. $r = 2$ 时称为是**均方相合**的.

在文献中, 相合一般即指弱相合.

上述定义也可类似地推广到一般的分布族 $\mathcal{P}$, 此时被估计的是总体分布的某个泛函 $\theta(P), P \in \mathcal{P}$.

上述几种收敛性质均为估计量最基本的大样本性质, 分别称为**相合性、强相合性和平均相合性**, 表示当样本量足够大时, 估计量可以在各自的意义下任意逼近要估计的量, 其重要性不言而喻. 例如, 一个估计量序列, 只有具备了(弱)相合性, 才能谈及极限分布等更深层次的渐近性质. 而在极限分布的证明过程中, 也往往需要首先建立估计序列的相合性.

由依概率收敛, a.s.收敛以及矩收敛之间的关系可知, 强相合性与平均相合性之间没有蕴含关系, 但均可导致弱相合性.

如果 $g(\theta) = \big(g_1(\theta), \cdots, g_k(\theta)\big)^\tau$ 是多维的, 则它的一列估计 $T_n = \big(T_{n1}, \cdots, T_{nk}\big)^\tau$ 称为在某种意义下相合, 若对 $i = 1, \cdots, k$, T_{ni} 是 $g_i(\theta)$ 在该意义下的相合估计. 因此在一般性讨论中只需要考虑一维情形的相合性. 容易证明如下定理.

定理 5.1 *设参数空间 Θ 为欧氏空间 $\mathbf{R}^k$ 的一个开子集, $g(\theta)$ 是 Θ 上的实值连续函数. 若 T_n 是 θ 的 (强或弱) 相合估计, 则 $g(T_n)$ 是 $g(\theta)$ 的 (强或弱) 相合估计.*

此定理不适用于平均相合性.

今设样本量为 n 时有统计量 T_n. 如果存在某个非退化分布 F, 以及参数 θ 的实值函数 $b_n(\theta) > 0$ 和 $a_n(\theta)$, 使得在任一 P_θ 之下, 当 $n \to \infty$ 时 $\big(T_n - a_n(\theta)\big)/b_n(\theta)$ 依分布收敛于 F, 则称 T_n 具有**渐近分布或极限分布**. 当 $F \sim N(0,1)$ 时, 称 T_n 具有**渐近正态性**. 当 T_n 为参数 $g(\theta)$ 的估计时, $a_n(\theta)$ 一般可取为 $g(\theta)$.

渐近正态的一个著名的例子为 i.i.d. 样本均值 $\bar{X}_n$. 设总体均值、方差分别为 $g(\theta)$ 和 $\sigma^2(\theta)>0$, 则依中心极限定理有 $\sqrt{n}\big(\bar{X}_n-g(\theta)\big)/\sigma(\theta)\xrightarrow{L}N(0,1)$. 因样本方差 S_n^2 为 $\sigma^2(\theta)$ 的一个相合估计, 用 S_n 代替 $\sigma(\theta)$, 即得 $\sqrt{n}\big(\bar{X}_n-g(\theta)\big)/S_n\xrightarrow{L}N(0,1)$. 渐近正态的这种形式在统计推断中用起来更方便.

考虑多维情形, 设 T_n 为 k 维统计量, $a_n(\theta)$ 为 k 维实向量, $B_n(\theta)>0$ 为 k 阶正定实方阵, 若当样本量 $n\to\infty$ 时 $B_n(\theta)^{-1}\big(T_n-a_n(\theta)\big)$ 依分布收敛于某 k 维非退化分布 F, 则称 T_n 具有**渐近分布或极限分布**. 当 F 为 k 维标准正态分布 $N(0,I_k)$ 时, 称 T_n 具有**渐近正态性**. 此处, I_k 为 k 阶单位阵.

下面引进**渐近有效估计** (Asymptotically Efficient Estimate, AEE) 的概念.

为行文简单计, 我们先限于考虑参数 θ 为一维的情形, 多维情形以及要估计的为参数函数的情形类似. 设 θ 的估计 $\hat{\theta}_n$ 有以下的渐近正态分布:

$$\sqrt{n}(\hat{\theta}_n-\theta)\xrightarrow{L}N\big(0,\sigma^2(\theta)\big),\quad n\to\infty. \tag{5.1}$$

此处, n 为样本量. 一般称 $\sigma^2(\theta)/n$ 为 $\hat{\theta}_n$ 的渐近方差. 显然, (5.1) 提供了一个在极限意义下比较估计优劣的准则: **渐近方差小者为优**. 这启示我们去考虑寻求渐近方差最小的估计. 今设分布族具有有限的 Fisher 信息量 $I(\theta)>0$, 我们从中抽得 i.i.d. 样本 X^n, 则在一定的条件下, 无偏估计的方差不能小于 Cramér-Rao 下界 $1/[nI(\theta)]$. 这启示我们这样想: 按 (5.1), 基于 X^n 的估计量 $\hat{\theta}_n$ 当 n 很大时可视为接近于无偏, 如果 C-R 不等式近似可用, 则 $\hat{\theta}_n$ 的渐近方差不能小于 $1/[nI(\theta)]$. 因此, 如果估计量满足

$$\sqrt{n}(\hat{\theta}_n-\theta)\xrightarrow{L}N\big(0,1/I(\theta)\big), \tag{5.2}$$

我们就可以把 $\hat{\theta}_n$ 叫做 θ 的**渐近有效估计**. 易见, 渐近有效估计如果存在, 必不唯一.

要使这个定义在理论上站得住脚, 需要解决两个问题:

1° $1/I(\theta)$ 确实是 (5.1) 右方极限方差的下界;

2° 这个下界可以达到, 即存在满足 (5.2) 的估计.

以后我们将证明, 在一定条件下, **极大似然估计**满足 (5.2). 至于第一个问题, 下面是一个反例.

例 5.1 (Hodges) 设 $X_1,\cdots,X_n$ 为抽自 $N(\theta,1)$ 的 i.i.d. 样本. 定义 θ 的估计 $\hat{\theta}_n$ 如下:

$$\hat{\theta}_n=\begin{cases}\bar{X}_n/2, & \text{若 } |\bar{X}_n|<n^{-1/4},\\ \bar{X}_n, & \text{若 } |\bar{X}_n|\geqslant n^{-1/4}.\end{cases}$$

容易证明, 当 $n\to\infty$ 时,

$$\sqrt{n}(\hat{\theta}_n-\theta)\xrightarrow{L}\begin{cases}N(0,1), & \text{若 } \theta\neq 0,\\ N(0,1/4), & \text{若 } \theta=0.\end{cases}$$

而 $I(\theta)\equiv 1$. 故当 $\theta=0$ 时上式的渐近方差突破了 $1/I(\theta)$ 这一下界.

这种 "**超有效点**" 的出现, 似乎使渐近有效估计的定义失去了意义. 但 Le Cam (1953) 证明: 对任意满足 (5.1) 的估计 $\hat{\theta}_n$, 超有效点集合仅为一 Lebesgue 零测集. 文

献中还指出, 如果要求 (5.1) 中的收敛对开区间中的参数 θ 有一致性, 则像例 5.1 这样的超有效点即不复存在. 因此, 渐近有效估计仍不失为一个站得住脚的概念.

有了上面的讨论, 我们不难对向量值的参数函数, 给出一列估计具有渐近有效性的一般性定义. 为此, 设 $X_1,\cdots,X_n$ 为取自分布族 $X\sim f(x,\theta)d\mu(x),\theta\in\Theta$ 的 i.i.d. 样本, μ 为样本空间 $(\mathcal{X},\mathcal{B})$ 上的一个 σ-有限测度, Θ 为 $\mathbf{R}^k$ 的一个开子集, $I(\theta):=\mathrm{E}_\theta[\partial\log f(X,\theta)/\partial\theta][\partial\log f(X,\theta)/\partial\theta^\tau]$ 为分布族的 Fisher 信息阵, $g:\Theta\mapsto\mathbf{R}^r$ 为一已知的可微函数, $T_n:r\times 1$ 为 $g(\theta)$ 的一列估计, 如果对任一 $\theta\in\Theta$ 在 P_θ 之下都有

$$\sqrt{n}\big(T_n-g(\theta)\big)\rightsquigarrow N_r\big(0,\partial g(\theta)/\partial\theta^\tau I^{-1}(\theta)\partial g^\tau(\theta)/\partial\theta\big),n\to\infty,$$

则称估计量 T_n **具有渐近有效性**, 或者称 T_n 为 $g(\theta)$ 的渐近有效估计. 这样的估计在有些文献中又称为**最优渐近正态估计** (Best Asymptotic Normal Estimates), 简记为 BAN 估计.

5.1.2 Delta 方法

我们常常会遇到这样的问题: 假定 k 维参数向量 θ 的估计 T_n 的渐近分布已经知道, 如何求出 $\phi(\theta)$ 的估计 $\phi(T_n)$ 的渐近分布? 此处, 映照 $\phi:D\subset\mathbf{R}^k\mapsto\mathbf{R}^m$ 在 θ 处可微, 而 θ 为 D 在 $\mathbf{R}^k$ 中的一个内点. 由可微性的定义,

$$\phi(\theta+h)-\phi(\theta)=\frac{\partial\phi(\theta)}{\partial\theta^\tau}h+o(||h||),\quad h\to 0. \tag{5.3}$$

此处, $\phi(\theta):=\big(\phi_1(\theta),\cdots,\phi_m(\theta)\big)^\tau,\theta=(\theta_1,\cdots,\theta_k)^\tau$,

$$\frac{\partial\phi(\theta)}{\partial\theta^\tau}=\begin{pmatrix}\partial\phi_1(\theta)/\partial\theta_1 & \cdots & \partial\phi_1(\theta)/\partial\theta_k\\ \partial\phi_2(\theta)/\partial\theta_1 & \cdots & \partial\phi_2(\theta)/\partial\theta_k\\ \vdots & & \vdots\\ \partial\phi_m(\theta)/\partial\theta_1 & \cdots & \partial\phi_m(\theta)/\partial\theta_k\end{pmatrix}.$$

将 $\phi(T_n)$ 在 θ 附近展开, 即可将其渐近分布的计算归结为 T_n 的渐近分布的计算. 这种方法在统计文献中称为 Delta 方法.

定理 5.2 设映照 $\phi:D\subset\mathbf{R}^k\mapsto\mathbf{R}^m$ 在 θ 处可微, θ 为 D 在 $\mathbf{R}^k$ 中的一个内点, T_n 为取值于 D 的 k 维随机向量. 假定对数列 $r_n\to\infty$, $r_n(T_n-\theta)\rightsquigarrow T,n\to\infty$, 则

$$r_n\big[\phi(T_n)-\phi(\theta)\big]=\frac{\partial\phi(\theta)}{\partial\theta^\tau}r_n(T_n-\theta)+o_p(1),$$

因而

$$r_n\big[\phi(T_n)-\phi(\theta)\big]\rightsquigarrow\frac{\partial\phi(\theta)}{\partial\theta^\tau}T.$$

证明 由 $r_n(T_n-\theta)$ 弱收敛知 $T_n-\theta$ 依概率收敛于 0, 因而 (5.3) 蕴含了

$$r_n\big(\phi(T_n)-\phi(\theta)\big)=\frac{\partial\phi(\theta)}{\partial\theta^\tau}r_n(T_n-\theta)+o_p(r_n||T_n-\theta||).$$

易见上式中的余项为 $o_p(1)$. 故由 Slutsky 定理, $r_n\big(\phi(T_n)-\phi(\theta)\big) \rightsquigarrow \big[\partial\phi(\theta)/\partial\theta^\tau\big]T$. ■

下面的定理是常用的, 它可由上一定理直接得出.

定理 5.3 *若在上述定理的条件下, 取 $r_n=\sqrt{n}$ 且假定 $\sqrt{n}(T_n-\theta)\rightsquigarrow N_k(0,\Sigma)$, $n\to\infty$, 则 $\sqrt{n}\big(\phi(T_n)-\phi(\theta)\big)\rightsquigarrow N_m\big(0,\partial\phi(\theta)/\partial\theta^\tau\Sigma\partial\phi^\tau(\theta)/\partial\theta\big)$.*

现设参数空间为一个一维开区间, θ 的估计量 T_n 满足 $\sqrt{n}(T_n-\theta)\overset{\theta}{\rightsquigarrow}N\big(0,\sigma^2(\theta)\big)$. 我们要据此构作 θ 的一个渐近置信区间. 因 T_n 的渐近方差与参数有关, 通常需要找到 $\sigma(\theta)$ 的一个相合估计 V_n. 此时基于极限分布 $\sqrt{n}(T_n-\theta)/V_n\overset{\theta}{\rightsquigarrow}N(0,1)$ 即可构作渐近置信区间. 另一种方法就是应用 Delta 方法, 找一个可微的严格单调函数 ϕ, 使 $\phi(T_n)$ 具有不依赖于参数 θ 的渐近方差. 这样一种变换在文献中称为**方差稳定变换**. 由定理 5.3, $\sqrt{n}\big(\phi(T_n)-\phi(\theta)\big)\rightsquigarrow N\big(0,\phi'(\theta)^2\sigma^2(\theta)\big)$, 因此只要取 $\phi(\theta)=\int d\theta/\sigma(\theta)$ 即可. 先构作 $\phi(\theta)$ 的一个渐近置信区间, 再将其转换为 θ 的渐近置信区间. 这样构作的渐近置信区间常导致更好的逼近.

例 5.2 设 $n\hat{p}_n\sim B(n,p), 0<p<1$. 当 $n\to\infty$ 时, $\sqrt{n}(\hat{p}_n-p)\rightsquigarrow N\big(0,p(1-p)\big)$. 变换

$$\phi(p)=\int dp/\sqrt{p(1-p)}=2\ \arcsin\sqrt{p}$$

即为一个方差稳定变换. 由此不难得出 p 的一个给定渐近水平 $1-\alpha$ 的渐近置信区间. 希望读者补出有关的证明和计算.

例 5.3 (相关系数) 设 $(X_1,Y_1),\cdots,(X_n,Y_n)$ 为取自相关系数为 ρ 的二维正态总体的 i.i.d. 样本. 样本相关系数定义为

$$r_n=\frac{\sum_{i=1}^n(X_i-\bar{X}_n)(Y_i-\bar{Y}_n)}{[\sum_{i=1}^n(X_i-\bar{X}_n)^2\sum_{i=1}^n(Y_i-\bar{Y}_n)^2]^{1/2}}.$$

应用 Delta 方法, 并通过繁冗的代数计算可得

$$\sqrt{n}(r_n-\rho)\rightsquigarrow N\big(0,(1-\rho^2)^2\big).$$

变换

$$\phi(\rho)=\int d\rho/(1-\rho^2)=(1/2)\log[(1+\rho)/(1-\rho)]=\operatorname{arctanh}\ \rho$$

即为一个方差稳定变换. 由此不难得出 ρ 的一个给定渐近水平 $1-\alpha$ 的渐近置信区间. 希望读者补出有关的计算.

5.1.3 矩估计

设 $X_1,\cdots,X_n$ 为随机元 X 的 i.i.d. 样本, $X\sim P_\theta$, $\theta:=(\theta_1,\cdots,\theta_k)^\tau\in\Theta$, 参数空间 Θ 为 $\mathbf{R}^k$ 的一个开子集. 若对给定的函数 $f_1,\cdots,f_k$, 通过解方程组

$$\frac{1}{n}\sum_{i=1}^n f_j(X_i)=\mathrm{E}_\theta[f_j(X)],\ j=1,\cdots,k,$$

可以得出 θ 的一个估计, 则所得估计 $\hat{\theta}_n$ 称为**矩估计**. 当函数 g 已知时, $g(\hat{\theta}_n)$ 称为 $g(\theta)$ 的矩估计. 若选取 $f_j(x)=x^j$, 则得经典的矩估计. 这种矩估计方法, 是现代统计学的奠基人之一 K. Pearson 在 20 世纪初研究曲线拟合时提出的, 至今仍不失为一种重要而常用的方法.

现设 X 为一维变量, r 为一正整数, 我们将总体的 r 阶原点矩和中心矩分别记为

$$\alpha_r(\theta)=\mathrm{E}_\theta(X^r),\quad \mu_r(\theta)=\mathrm{E}_\theta[X-\alpha_1(\theta)]^r,$$

而将 r 阶样本原点矩和中心矩分别记为

$$a_{rn}=\frac{1}{n}\sum_{i=1}^n X_i^r,\quad m_{rn}=\frac{1}{n}\sum_{i=1}^n (X_i-\bar{X}_n)^r.$$

易见, 样本均值 $\bar{X}_n$ 就是一阶样本原点矩. 在经典的矩估计方法中, 涉及的即为总体和样本的各阶原点矩和样本矩构成的方程组, 其中的总体矩是作为 θ 的函数出现的.

从不同的矩出发估计某个参数, 有时会得到不同的估计. 例如, Poisson 分布参数 θ 既是总体均值又是总体方差, 按矩估计法, 既可用样本均值 a_{1n}, 也可用 2 阶样本中心矩 m_{2n} 去估计它. 因此, 矩估计法应视为一个原则而不是确定的公式. 一般讲, 在类似情形, 在经典的矩估计中, 应优先考虑使用阶数较低的矩.

引进向量记号 $f=(f_1,\cdots,f_k)^\tau$, 记 $P_\theta f:=\mathrm{E}_\theta[f(X)]$, $\mathbb{P}_n f=n^{-1}\sum_{i=1}^n f(X_i)$, 且用 $\psi:\Theta\mapsto\mathbf{R}^k$ 记由 $\psi(\theta):=P_\theta f$ 定义的变换, 则矩估计 $\hat{\theta}_n$ 由向量方程

$$\mathbb{P}_n f=\psi(\theta)\quad(\equiv P_\theta f) \tag{5.4}$$

的解给出. 为了给出矩估计的存在性定理并研究它的渐近性质, 我们引进下面的定理.

定理 5.4 (逆映照存在定理) *设 A 为 $\mathbf{R}^k$ 的一个开子集, 映照 $g:A\mapsto\mathbf{R}^k$ 在 $a\in A$ 的一个邻域中可微, 在 a 点连续可微, 且 Jacobi 矩阵 $Dg(x):=\partial g(x)/\partial x^\tau$ 在 a 点非奇异, 则存在 a 的一个开邻域 U, g 将 U 一对一地映为 $\mathbf{R}^k$ 的一个开子集 V. 逆映照 $g^{-1}:V\mapsto U$ 存在、连续且在 V 中可微, 当 $y=g(x)$ 时有 $Dg^{-1}(y):=\partial g^{-1}(y)/\partial y^\tau=\big(Dg(x)\big)^{-1}$.*

证明可参看 Munkres (1991). 书中的证明也适用于此处较弱的条件.

定理 5.5 *设参数空间 Θ 为 $\mathbf{R}^k$ 的一个开子集, 映照 $\psi:\Theta\mapsto\mathbf{R}^n$ 在 Θ 上连续可微, 且 $D\psi(\theta):=\partial\psi(\theta)/\partial\theta^\tau$ 在 Θ 上非奇异, 则在任一 P_θ 之下, 以概率为 1 当 n 充分大时, 矩估计方程 (5.4) 在 θ 的一个邻域中存在唯一解 $\hat{\theta}_n$, 这样得到的 $\hat{\theta}_n$ 具有强相合性, 且若 $P_\theta||f||^2<\infty$, 则*

$$\sqrt{n}(\hat{\theta}_n-\theta)\overset{\theta}{\rightsquigarrow}N_k\big(0,[\partial\psi(\theta)/\partial\theta^\tau]^{-1}[P_\theta f f^\tau-P_\theta f P_\theta f^\tau][\partial\psi^\tau(\theta)/\partial\theta]^{-1}\big).$$

此处, $\psi(\theta)=P_\theta f$.

证明 任取 $\theta^0\in\Theta$, 由逆映照存在定理, 存在 θ^0 的一个开邻域 U, ψ 将 U 一对一地映为 $\psi(\theta^0)=P_{\theta^0}f$ 在 $\mathbf{R}^k$ 中的一个开邻域 V, 因而可以定义 ψ 的逆映照 $\psi^{-1}:V\mapsto U$. 由强大数律, 在 P_{θ^0} 之下, $\mathbb{P}_n f\to P_{\theta^0}f$, a.s., 故以概率 1, 当 n 充分大

时, $\mathbb{P}_n f \in V$, 此时矩估计方程 (5.4) 在 U 中存在唯一解 $\hat{\theta}_n = \psi^{-1}(\mathbb{P}_n f)$, 且 $\hat{\theta}_n$ 具有强相合性. 中心极限定理保证了 $\sqrt{n}(\mathbb{P}_n f - P_{\theta^0} f)$ 的渐近正态性, 再由 Delta 方法及逆映照微分法则, 即可建立 $\sqrt{n}(\hat{\theta}_n - \theta^0) = \sqrt{n}\big(\psi^{-1}(\mathbb{P}_n f) - \psi^{-1}(P_{\theta^0} f)\big)$ 的渐近正态性. ■

关于矩估计的平均相合性, 我们将仅限于证明一个关于样本均值 r 阶矩收敛的结果. 通过下面的例子, 细心的读者可以领略一下如何精确估计收敛的阶. 为此, 我们先引进两个预备定理. 这两个预备定理在我们以后的研究工作中还会经常用到.

定理 5.6 设 $X_1, \cdots, X_n$ 为相互独立、r 阶矩有限且均值为 0 的随机变量, $1 < r < \infty$, 则存在仅依赖于 r 的常数 C_1, C_2, 使

$$C_1 \mathrm{E}\Big|\sum_{i=1}^n X_i^2\Big|^{r/2} \leqslant \mathrm{E}\Big|\sum_{i=1}^n X_i\Big|^r \leqslant C_2 \mathrm{E}\Big|\sum_{i=1}^n X_i^2\Big|^{r/2}.$$

上述不等式称为 Marcinkiewicz 不等式, 它同样适用于鞅差序列, 称为 Marcinkiewicz-Zygmund-Burkholder 不等式, 其证明可参看 Hall 和 Heyde (1980).

引理 5.7 (Kronecker 引理) 设级数 $\sum_{n=1}^\infty x_n$ 收敛, $0 < b_n \uparrow \infty$, 则

$$\frac{1}{b_n}\sum_{k=1}^n b_k x_k \to 0, \quad n \to \infty.$$

例 5.4 设 $X_1, \cdots, X_n$ 为 X 的 i.i.d. 样本, $\bar{X}_n$ 为样本均值, 且 $\mathrm{E}\,|X|^r < \infty, r \geqslant 1$ 不必为整数, 则

$$\mathrm{E}\,|\bar{X}_n - \alpha_1|^r = \begin{cases} o(n^{1-r}), & 1 \leqslant r < 2, \\ O(n^{-r/2}), & r \geqslant 2. \end{cases} \tag{5.5}$$

为证 (5.5), 不妨设 $\alpha_1 := \mathrm{E}(X) = 0$. 下面用 C 来表示常数, 每次出现允许取不同的值.

当 $r \geqslant 2$ 时, 由 Marcinkiewicz 不等式及 Jensen 不等式得出

$$\begin{aligned} \mathrm{E}\,|\bar{X}_n|^r &\leqslant Cn^{-r}\mathrm{E}\Big|\sum_{i=1}^n X_i^2\Big|^{r/2} \\ &\leqslant Cn^{-r/2}\mathrm{E}\Big[\sum_{i=1}^n |X_i|^r/n\Big] = Cn^{-r/2}\mathrm{E}\,|X|^r. \end{aligned}$$

这就证明了 (5.5) 中 $r \geqslant 2$ 的部分.

为了在 $1 \leqslant r < 2$ 情形得到所要的阶, 下面采用截尾和中心化方法. 为此, 记

$$\begin{aligned} &X_i' = X_i I(|X_i| > n^{1/r}), \quad X_i'' = X_i I(|X_i| \leqslant n^{1/r}), \\ &S_n' = \sum_{i=1}^n X_i', \qquad\qquad S_n'' = \sum_{i=1}^n X_i'', \\ &T_n' = \sum_{i=1}^n (X_i' - \mathrm{E}\,X_i'), \quad T_n'' = \sum_{i=1}^n \big(X_i'' - \mathrm{E}\,X_i''\big). \end{aligned}$$

(i) $r = 1$ 时由 $\mathrm{E}\,|X| < \infty$ 及控制收敛定理, 得

$$\mathrm{E}\,|S_n'/n| \leqslant \mathrm{E}\,|X| I(|X| > n) = o(1).$$

(ii) $1 \leqslant r < 2$ 时, 由 $\mathrm{E}\,|X|I(|X| > n^{1/r}) \leqslant n^{-(r-1)/r}\mathrm{E}\,|X|^r I(|X|^r > n)$ 及 $\mathrm{E}(X_i) = 0$, 可得

$$|\mathrm{E}(S_n''/n)|^r = |\mathrm{E}(S_n'/n)|^r = o(n^{1-r}).$$

(iii) $1 < r < 2$ 时, 由 Marcinkiewicz 不等式及 $0 < r/2 < 1$, 有

$$\begin{aligned}\mathrm{E}\,|T_n'/n|^r &\leqslant Cn^{-r}\mathrm{E}\Big|\sum_{i=1}^n (X_i' - \mathrm{E}\,X_i')^2\Big|^{r/2}\\ &\leqslant Cn^{-r}\mathrm{E}\Big[\sum_{i=1}^n |X_i' - \mathrm{E}\,X_i'|^r\Big] = o(n^{1-r}).\end{aligned}$$

(iv) $1 \leqslant r < 2$ 时, 由 Jensen 不等式,

$$\begin{aligned}\mathrm{E}\,|T_n''/n|^r &\leqslant n^{-r}(\mathrm{E}\,|T_n''|^2)^{r/2} = n^{-r}[n\mathrm{E}\,(X_1'' - \mathrm{E}\,X_1'')^2]^{r/2}\\ &\leqslant n^{-r/2}[\mathrm{E}(X_1'')^2]^{r/2} = n^{-r/2}[\mathrm{E}\,X^2 I(|X| \leqslant n^{1/r})]^{r/2}\\ &\leqslant n^{-r/2}\Big\{\sum_{k=1}^n k^{2/r}P(k-1 < |X|^r \leqslant k)\Big\}^{r/2}.\end{aligned}$$

因 $\mathrm{E}\,|X|^r < \infty$ 当且仅当 $\sum_{k=1}^\infty kP(k-1 < |X|^r \leqslant k) < \infty$, 在引理 5.7 中取 $b_k = k^{2/r-1}$, 得

$$\frac{1}{n^{2/r-1}}\sum_{k=1}^n k^{2/r}P(k-1 < |X|^r \leqslant k) \to 0,$$

因此, $\mathrm{E}\,|T_n''/n|^r = o(n^{1-r})$.

由 (i) 到 (iv) 的有关结果及 C_r 不等式, 即可证得当 $1 \leqslant r < 2$ 时 (5.5) 成立.

上述结果已经推广到 U 统计量, 参看陈希孺 (1980).

5.2 极大似然估计

似然方法是广泛使用的一种统计方法, 虽可将其归于 M-方法这一庞大的家族, 但它具有独立的意义, 在估计和检验问题中都有广泛的应用. 随着统计研究的不断深入和发展, 文献中已经提出了多种似然方法, 除了通常的似然和似然比以外, 还有条件似然、偏似然、剖面似然、拟似然、经验似然等多种. 本书将只限于讨论通常用于参数估计的极大似然方法和用于假设检验的似然比方法.

5.2.1 一般概念

设 $\{f(x,\theta)d\mu(x), \theta \in \Theta\}$ 为样本分布族, μ 为样本空间 $(\mathcal{X}, \mathcal{B})$ 上的一个 σ-有限测度. 前已指出, 设有了样本 x, 若将 $f(x,\cdot)$ 视为参数空间 Θ 上的函数, 则称为 (样本 x 的) **似然函数**. 似然函数若在 Θ 上的某点 $\hat\theta(x)$ 取得最大值, 即

$$f\big(x, \hat\theta(x)\big) = \max\{f(x,\theta) : \theta \in \Theta\},$$

则称 $\hat{\theta}(x)$ 为未知参数 θ 的**极大似然估计** (Maximum Likelihood Estimate), 简记为 MLE. 若要估计的是某个参数函数 $g(\theta)$, 则 $g\big(\hat{\theta}(x)\big)$ 称为 $g(\theta)$ 的 MLE.

提到 MLE, 就不能不提到 Fisher 的有关工作. 在他 1912 年的一项工作中, 极大似然方法首次被用于未知参数的估计, 这是他对统计学最早的贡献. 在接下来的十二三年中, 他发表了一系列论文 (如 1922, 1925), 为现代统计学成为一门独立而具有巨大应用潜力的科学分支奠定了理论基础. 在他早期关于估计理论的工作中, MLE 处在很重要的位置. 虽然估计的似然方法在 Gauss (1821), Laplace (1820), Edgeworth (1908, 1909) 甚至更早的文献中出现过, 但正是 Fisher 看到了 MLE 广泛用于估计问题的潜在作用, 把它作为估计的一般方法加以推荐, 并开始了对 MLE 性质的研究. Fisher 的有关工作具有先驱的性质. 继 Fisher 的工作之后, 产生了对 MLE 的相合性和渐近效率的乐观信念, 并获得了一系列重要的理论结果, 例如 Cramér (1946a) (1946b), Wald (1949), Le Cam (1953) 等.

以后我们会讲到, 在某些 "规则" 条件下, 极大似然方法给出了最优渐近正态的估计量. 当然, 在统计中没有一种已知的方法是普遍最优的, 不同的方法可能适用于不同的目的和不同的场合, 极大似然方法也不例外. 当面对新的问题需要寻求新的估计时, 极大似然方法仍然不失为人们的首选.

关于 Fisher 对现代统计学的贡献, 可参看 Rao (1992).

今设样本 $X \sim f(x,\theta)d\mu(x), \theta = (\vartheta_1, \cdots, \vartheta_k)^\tau$, 参数空间 Θ 为 $\mathbf{R}^k$ 中一个有内点的子集, 且似然函数 $f(x,\cdot)$ 在 Θ 的内部存在各个偏导数, 则方程(组)

$$\partial f(x,\theta)/\partial\vartheta_j = 0,\ j = 1, \cdots, k \tag{5.6}$$

称为**似然方程**. 如果 $f(x,\cdot)$ 的最大值在 Θ 的某个内点处达到, 则似然方程必有一解就是 MLE. 但在方程有多个解的情形, 要确定哪个解是最大值点, 可能并不容易.

现设 $X_1, \cdots, X_n$ 为取自分布 $f(x,\theta)d\mu(x)$ 的 i.i.d. 样本, 简记为 X^n. 这种情形不难归结到上述定义的架构之中, 此时似然函数为 $L = \Pi_{i=1}^n f(x_i,\theta)$. 若借助于**对数似然函数** $\log L$, 则可将似然方程写成如下更为方便的形式:

$$\partial \log L/\partial\vartheta_j \equiv \sum_{i=1}^n \partial \log f(x_i,\theta)/\partial\vartheta_j = 0, \quad j = 1, \cdots, k, \tag{5.7}$$

上述方程也可称为**对数似然方程**.

例 5.5 似然方法不仅适用于独立样本, 也适用于相依序列的观测值. 下面就是 Markov 链的一个简单的例子.

设 $X_1 \sim N(0,\sigma^2)$, 而当给定 $X_1, \cdots, X_{i-1}$ 时, X_i 的条件分布只与 X_{i-1} 有关, 这种性质称为 Markov 性. 今设 $X_i\big|X_{i-1} = x_{i-1} \sim N(\alpha x_{i-1} + \beta, \sigma^2)$, 而观测值为 $(X_1, \cdots, X_n)$, 则其联合分布为

$$\Big(\frac{1}{2\pi\sigma^2}\Big)^{n/2} \exp\Big\{-\frac{1}{2\sigma^2}\Big[x_1^2 + \sum_{i=2}^n (x_i - \alpha x_{i-1} - \beta)^2\Big]\Big\} dx_1 \cdots dx_n.$$

如不计一个与参数无关的常数, 对数似然函数为

$$\log L(\alpha,\beta,\sigma^2|x^n)=-\frac{n}{2}\log\sigma^2-\frac{1}{2\sigma^2}\Big[x_1^2+\sum_{i=2}^{n}(x_i-\alpha x_{i-1}-\beta)^2\Big].$$

由此即得似然方程组

$$\begin{cases}\sum\limits_{i=2}^{n}x_{i-1}(x_i-\alpha x_{i-1}-\beta)=0,\\ \sum\limits_{i=2}^{n}(x_i-\alpha x_{i-1}-\beta)=0,\\ n/(2\sigma^2)-[1/(2\sigma^4)]\{x_1^2+\sum_{i=2}^{n}(x_i-\alpha x_{i-1}-\beta)^2\}=0,\end{cases}$$

即

$$\begin{cases}(\sum\limits_{i=2}^{n}x_{i-1}^2)\alpha+(\sum\limits_{i=2}^{n}x_{i-1})\beta=\sum\limits_{i=2}^{n}x_{i-1}x_i,\\ (\sum\limits_{i=2}^{n}x_{i-1})\alpha+(n-1)\beta=\sum\limits_{i=2}^{n}x_i,\\ \sigma^2=(1/n)\{x_1^2+\sum\limits_{i=2}^{n}(x_i-\alpha x_{i-1}-\beta)^2\}.\end{cases}$$

由该方程组解出的 $(\hat\alpha,\hat\beta,\hat\sigma^2)$ 即为 (α,β,σ^2) 的MLE.

例 5.6 当似然函数在某些点不连续或不可导时, 寻求 MLE 已不能求助于似然方程, 只能具体情况具体分析. 下面的例子来自一维截断分布族, 式中的 h 为定义于 $(-\infty,\infty)$ 而恒取正值的一个函数.

(i) 单边截断分布族:

$$f(x,\theta)dx=C(\theta)h(x)I(0<x<\theta)dx,\quad 0<\theta<\infty,\tag{5.8}$$

其中 $C(\theta)=1/\int_0^\theta h(x)dx$.

(ii) 双边截断分布族:

$$f(x,\theta_1,\theta_2)dx=C(\theta_1,\theta_2)h(x)I(\theta_1<x<\theta_2)dx,\quad -\infty<\theta_1<\theta_2<\infty,\tag{5.9}$$

其中 $C(\theta_1,\theta_2)=1/\int_{\theta_1}^{\theta_2}h(x)dx$.

设 $X_1,\cdots,X_n$ 为取自分布族 (i) 或 (ii) 的 i.i.d. 样本. 记

$$X_{(1)}:=\min(X_1,\cdots,X_n),\quad X_{(n)}:=\max(X_1,\cdots,X_n).$$

对单边情形, 当 θ 减少时 $C(\theta)$ 严增, 若将似然函数写为

$$L=C(\theta)^n[\prod_{i=1}^{n}h(x_i)I(x_i>0)]I(x_{(n)}\leqslant\theta),$$

则为要 $L\neq0$, 应有 $\theta\geqslant x_{(n)}$, 此时 θ 取值越小, L 取值越大, 故 θ 的 MLE 为 $X_{(n)}$.

对双边情形, 当 θ_1 增加及 θ_2 减少时, $C(\theta_1,\theta_2)$ 严格增加. 故若将似然函数写成

$$L=C(\theta_1,\theta_2)^n[\prod_{i=1}^{n}h(x_i)]I(\theta_1\leqslant x_{(1)}\leqslant x_{(n)}\leqslant\theta_2),$$

知 θ_1 和 θ_2 的 MLE 分别为 $X_{(1)}$ 和 $X_{(n)}$.

5.2.2 指数族情形的 MLE

设样本 X 的分布属于自然形式的指数族

$$dP_\theta(x) = f(x,\theta)d\mu(x) = C(\theta)\exp\{\theta^\tau T(x)\}d\mu(x), \quad \theta \in \Theta, \tag{5.10}$$

此处, $T(x) = (T_1(x), \cdots, T_k(x))^\tau$, $\Theta \subset \mathbf{R}^k$ 为自然参数空间.

此处我们作如下假定 (参看 1.2.2 节注 1.2):

(i) 假定 Θ 在 $\mathbf{R}^k$ 中有内点, 并将 Θ 的内部记为 Θ^{o};

(ii) 假定对任何 $\theta \in \Theta^{\mathrm{o}}$, 不存在非零的常数向量 a, 使得线性组合 $a^\tau T(X)$ 在 P_θ 之下退化为一个常数, 或者等价地, 即假定 $T(X)$ 在 P_θ 之下的协方差阵 $\mathrm{Cov}_\theta[T(X)]$ 为非奇异的.

通常称满足这样两个条件的指数族是满秩的.

定理 5.8 *假定在分布族 (5.10) 之下, 条件 (i) 和 (ii) 成立, 则似然方程在 Θ^{o} 中至多有一个解, 且如有一解, 此解就是唯一的 MLE.*

证明 由指数族的性质和满秩的假定, Θ^{o} 为 $\mathbf{R}^k$ 中的一个开凸集, 且在 Θ^{o} 上, 矩阵函数 $\partial^2 \log C(\theta)/\partial\theta\partial\theta^\tau = -\mathrm{Cov}_\theta[T(X)]$ 恒为负定. 故对数似然函数 $l(\theta) = \log f(x,\theta) = \log C(\theta) + \theta^\tau T(x)$ 对 θ 严格凹. 这种严格凹的性质, 决定了对数似然方程在 Θ^{o} 内至多有一个解. 直接证明也很容易: 如若不然, 假定 $\partial l(\theta)/\partial\theta$ 在 Θ^{o} 内有两个零点 $\zeta_0 \neq \zeta_1$, 记 $\zeta_t = (1-t)\zeta_0 + t\zeta_1, g(t) = l(\zeta_t), 0 \leqslant t \leqslant 1$, 则函数 g 有任意阶导数, 且 $g'(0) = g'(1) = 0$, 由中值定理, 存在 $s \in (0,1)$, 使 $0 = g''(s) = (\zeta_1 - \zeta_0)^\tau \partial^2 l(\zeta_s)/\partial\theta\partial\theta^\tau (\zeta_1 - \zeta_0) < 0$. 这个矛盾即证明了上述论断. 若设似然方程有一个解 $\xi = \xi(x)$, 则 ξ 必为似然函数在 Θ^{o} 上的最大值点.

今设 η 为 $\Theta - \Theta^{\mathrm{o}}$ 中的任意一点, 为证 ξ 为唯一的 MLE, 只需要证明 $f(x,\eta) < f(x,\xi)$. 为此, 令 θ_n 沿着线段 $\overline{\xi\eta}$ 趋于 η, $\theta_n = a_n\xi + b_n\eta \in \Theta^{\mathrm{o}}, a_n > 0, b_n = 1 - a_n > 0$, 且 $a_n \downarrow 0$. 限制在 $\overline{\xi\eta} \cap \Theta^{\mathrm{o}}$ 上, $l(\theta)$ 是一维的严格凹函数, 故对每个 θ_n, 有 $f(x,\xi) > f(x,\theta_n)$, 且 $f(x,\theta_n)$ 为下降序列. 易见 $\exp\{\theta_n^\tau T\} \to \exp\{\eta^\tau T\}$. 由自然参数空间的定义, 应有

$$0 < \int \exp\{\xi^\tau T\}d\mu < \infty, \quad 0 < \int \exp\{\eta^\tau T\}d\mu < \infty,$$

且 $\exp\{\theta_n^\tau T\} \leqslant \exp\{\xi^\tau T\} + \exp\{\eta^\tau T\}$; 故由控制收敛定理, 有 $\int \exp\{\theta_n^\tau T\}d\mu \to \int \exp\{\eta^\tau T\}d\mu$, $C(\theta_n) \to C(\eta)$, $f(x,\theta_n) \downarrow f(x,\eta)$. 因而 $f(x,\eta) < f(x,\xi)$. 这就证明了 ξ 为唯一的 MLE. ■

指数族参数的 MLE 的大样本性质

假定自然形式的指数族 (5.10) 是满秩的, $X_1, \cdots, X_n$ 为从该分布族中抽取的 i.i.d. 样本, 为简单计, 不妨假定参数空间 Θ 仅限于自然参数空间的内部, 它是 $\mathbf{R}^k$ 中的一个开凸集. 记 $T(X) = T, T(X_i) = T_i, \bar{T}_n = n^{-1}\sum_{i=1}^n T_i$, 则对数似然函数为

$$\log L(\theta|X_1, \cdots, X_n) = n\log C(\theta) + n\theta^\tau \bar{T}_n. \tag{5.11}$$

对数似然方程为

$$\psi(\theta) := -\partial \log C(\theta)/\partial\theta = \bar{T}_n. \tag{5.12}$$

由此可见, 似然方程和矩估计方程 $\mathrm{E}_\theta(T) = \bar{T}_n$ 是一致的.

由指数族满秩的假定, 映照 $\psi : \Theta \mapsto \mathbf{R}^k$ 的 Jacobi 矩阵

$$D\psi(\theta) = -\frac{\partial^2 \log C(\theta)}{\partial\theta\partial\theta^\tau} = \mathrm{Cov}_\theta[T(X)]$$

在 Θ 上恒为正定. 结合定理 5.8 和定理 5.5, 在任一 P_θ 之下, 以概率 1 当 n 充分大时, 似然方程 (5.12) 在 Θ 中有唯一解 $\hat{\theta}_n = \psi^{-1}(\bar{T}_n)$, $\hat{\theta}_n$ 是强相合的, 且对任一 $\theta \in \Theta$, 有

$$\sqrt{n}(\hat{\theta}_n - \theta) \overset{\theta}{\rightsquigarrow} N_k\big(0, I(\theta)^{-1}\big), \tag{5.13}$$

其中, $I(\theta) = D\psi(\theta) = \mathrm{Cov}_\theta[T(X)]$ 为该指数族的 Fisher 信息矩阵.

以上结果不难推广到指数族

$$f_\theta(x)d\mu := f(x,\theta)d\mu = C^*(\theta)\exp\{Q(\theta)^\tau T(x)\}d\mu, \quad \theta \in \Theta \tag{5.14}$$

的一般情形, 此处 $T(x) = (T_1(x), \cdots, T_k(x))^\tau$, $Q(\theta) = \big(Q_1(\theta), \cdots, Q_k(\theta)\big)^\tau$. 假定 Θ 为 $\mathbf{R}^k$ 的一个开子集, $Q : \Theta \mapsto \mathbf{R}^k$ 是一对一的连续可微的变换, 其偏导数阵 $D(\theta) := \partial Q(\theta)/\partial\theta^\tau$ 非奇异, 且指数族 (5.14) 满秩. 记 $\dot{l}_\theta = \partial \log f(x,\theta)/\partial\theta$, 则对任一 $\theta \in \Theta$, 在 P_θ 之下, 以概率 1 当 n 充分大时, 似然方程有唯一解 $\hat{\theta}_n$, 且

$$\sqrt{n}(\hat{\theta}_n - \theta) \rightsquigarrow N_k(0, I_\theta^{-1}), \quad I_\theta = P_\theta \dot{l}_\theta \dot{l}_\theta^\tau = D(\theta)^\tau \mathrm{Cov}_\theta\big(T(X)\big)D(\theta).$$

证明留给读者作为练习.

例 5.7 即使样本服从自然形式的指数族分布, MLE 也可能不存在. 例如, 若取样本 X 的分布为

$$P_\theta(X=1) = 1 - P_\theta(X=0) = \mathrm{e}^\theta/(1+\mathrm{e}^\theta), \ -\infty < \theta < \infty,$$

则当 $x=1$ 或 $x=0$ 时, 似然函数分别为 $\mathrm{e}^\theta/(1+\mathrm{e}^\theta)$ 和 $1/(1+\mathrm{e}^\theta)$, 由于各自的严格单调性, 两者在 $(-\infty,\infty)$ 上均无极值点.

现设 $X_1, \cdots, X_n$ 为从上述分布族中抽取的 i.i.d. 样本, 要寻求 θ 的 MLE $\hat{\theta}_n$. 似然函数为

$$\Big(\frac{\mathrm{e}^\theta}{1+\mathrm{e}^\theta}\Big)^{n\bar{X}_n}\Big(\frac{1}{1+\mathrm{e}^\theta}\Big)^{n(1-\bar{X}_n)}, \ -\infty < \theta < \infty.$$

似然方程为

$$\bar{X}_n = \mathrm{e}^\theta/(1+\mathrm{e}^\theta).$$

当 $\bar{X}_n = 0$ 或 $\bar{X}_n = 1$ 时, 似然方程无解同时 MLE 不存在; 而当 $0 < \bar{X}_n < 1$ 时, 方程有解 $\hat{\theta}_n = \log[\bar{X}_n/(1-\bar{X}_n)]$ 且此解即为 MLE. 而在任一 P_θ 之下, MLE 不存在的概率 $\big(\mathrm{e}^\theta/(1+\mathrm{e}^\theta)\big)^n + \big(1/(1+\mathrm{e}^\theta)\big)^n \to 0, n \to \infty$, 但在固定样本情形, 这个概率总是正的. 这跟指数族的一般理论是一致的.

5.2.3 MLE 的渐近理论

关于 MLE 的渐近理论, 可以从两个角度讨论. 一是从 MLE 的极值定义出发, 研究似然函数整体最大值点的极限性质. 用这种方式处理难度较大. 另一种则是遵循 Cramér (1946a) (1946b) 开创的路线, 将重点从整体最大转移到局部极大, 并在一组 "规则" 条件下建立似然方程相合解的渐近有效性. 沿着这条路径处理起来相对容易一些. 在本节中, 我们将遵循后一条路线, 讨论似然方程解的渐近理论, 包括解的存在性、可测性、相合性和渐近有效性.

1. 似然方程解的存在性、可测性和相合性

我们首先讨论解的存在性、可测性和相合性, 先从一维参数的情形谈起. 为简单计, 下面将样本 $X_1,\cdots,X_n$ 简记为 X^n. 我们有下述定理.

定理 5.9 *设 $X\sim dP_\theta(x)=f(x,\theta)d\mu(x)$, $X_1,\cdots,X_n$ 为从中抽取的 i.i.d. 样本, 参数空间 Θ 为一开区间. 假定:*

(i) *若 $\theta_1\neq\theta_2$, 则 $P_{\theta_1}\neq P_{\theta_2}$, 即不同的参数值对应于不同的分布.*

(ii) *对 X 值域空间中任给的 x, $f(x,\theta)$ 在 Θ 上处处大于 0 且对 θ 有一阶连续偏导数.*

那么在任一 P_θ 之下, 存在一列随机变量 $\{\hat\theta_n\}$, 以概率 1 当 n 充分大时满足似然方程

$$\frac{\partial\log L(X^n,\theta)}{\partial\theta}:=\sum_{i=1}^n\frac{\partial\log f(X_i,\theta)}{\partial\theta}=0, \tag{5.15}$$

且是强相合的.

定理结论有时也叙述为: 在任一 P_θ 之下, 以概率 1 当 n 充分大时, 似然方程 (5.15) 存在一个可测解, 且解序列 $\{\hat\theta_n\}$是强相合的.

证明 任取 $\theta_0\in\Theta$, 并取 $a>0$ 充分小, 使区间 $B_a\equiv[\theta_0-a,\theta_0+a]\subset\Theta$. 由条件 (i) 及引理 1.28, 有 $\mathrm{E}_{\theta_0}\log[f(X,\theta_0\pm a)/f(X,\theta_0)]<0$. 由强大数定律, 在 P_{θ_0} 之下, 以概率 1, 当 n 充分大时有 $\sum_{i=1}^n\log[f(X_i,\theta_0\pm a)/f(X_i,\theta_0)]<0$, 即

$$\sum_{i=1}^n\log f(X_i,\theta_0\pm a)<\sum_{i=1}^n\log f(X_i,\theta_0). \tag{5.16}$$

故在区间 B_a 内部, 似然函数有一局部极大, 似然方程 (5.15) 在其中有解. 由函数 $\partial\log L(X^n,\theta)/\partial\theta$ 的连续性, 知其零点的极限仍为其零点, 因而

$$\hat\theta_{n,a}:=\inf\{\theta\geqslant\theta_0-a:\partial\log L(X^n,\theta)/\partial\theta=0\}$$

必为似然方程的一个根.

定义 Ω_a 为这样的样本序列 $\omega:=(X_1,X_2,\cdots)$ 的集合, 对这些 ω, 存在 $N_a(\omega)$, 当 $n\geqslant N_a(\omega)$ 时, 似然方程 (5.15) 存在上述解 $\hat\theta_{n,a}:=\hat\theta_{n,a}(\omega)$. 由上面的论证, $P_{\theta_0}(\Omega_a)=1$.

今证 $\hat{\theta}_{n,a}:\mathcal{X}^n\mapsto\Theta$ 为 Borel 可测. 为此, 设 $t\geqslant\theta_0-a$ 并用 Q 记有理数集, 则由似然函数的连续性和 $\hat{\theta}_{n,a}$ 的定义,

$$\{\hat{\theta}_{n,a}>t\}:=\{x^n:\hat{\theta}_{n,a}>t\}=\Big\{\inf_{\theta_0-a\leqslant\theta\leqslant t,\theta\in Q}\Big|\frac{\partial\log L(x^n,\theta)}{\partial\theta}\Big|>0\Big\},\tag{5.17}$$

故 $\hat{\theta}_{n,a}$ 为 Borel 可测.

令 $\Omega_0=\bigcap_{k=1}^{\infty}\Omega_{a/k}$, 则 $P_{\theta_0}(\Omega_0)=1$. 不妨设 $N_a(\omega)\leqslant N_{a/2}(\omega)\leqslant N_{a/3}(\omega)\leqslant\cdots$. 当 $N_{a/k}(\omega)\leqslant n<N_{a/(k+1)}(\omega)$ 时, 定义 $\hat{\theta}_n(\omega)=\hat{\theta}_{n,a/k}(\omega)$; 而当 $n<N_a(\omega)$ 或 $\omega\notin\Omega_0$ 时, 定义 $\hat{\theta}_n(\omega)$ 为 Θ 内的任一常数, 则 $\{\hat{\theta}_n\}$ 是一列随机变量, 在 P_{θ_0} 之下, 以概率 1 当 n 充分大时满足似然方程, 且由构造方式知其是强相合的. ■

此处的证明参考了 Serfling (1980). 值得注意的是, 在有关可测性的证明中, 我们用到了 f 对 θ 的偏导数的连续性. 如果在上述证明中定义 $\hat{\theta}_n(a)$ 为似然方程在区间 $[\theta_0-a,\theta_0+a]$ 上与 θ_0 最近的零点, 同样可以证明, 对任给的 t, $\{\hat{\theta}_n(a)>t\}$ 为一可测集. 由此即可用类似方法构造一列似然方程的强相合的可测解. 我们把这一结论的证明留给读者作为练习.

注 5.1 (a) 要在普适的条件下, 建立 MLE 以及似然方程的解的可测性, 一般并不容易. 今后我们不再予以讨论.

(b) 定理 5.9 证明中涉及的 $\hat{\theta}_n$, 因其依赖于未知参数, 还不是一个估计量.

现设除定理条件之外, 还存在参数的一串强 (弱) 相合估计 $\{\bar{\theta}_n\}$, 则存在 θ 的一串强 (弱) 相合估计, 以概率 1 当 n 充分大时为似然方程的根.

事实上, 因似然方程所有的根构成 $\mathbf{R}^k$ 中的一个闭集, 故可从中选出一个最接近于 $\bar{\theta}_n$ 的, 记为 θ_n^*, 则 θ_n^* 确是与参数无关的估计量, 且由定理 5.9, 以及不等式 $|\theta_n^*-\theta_0|\leqslant|\theta_n^*-\bar{\theta}_n|+|\bar{\theta}_n-\theta_0|\leqslant|\hat{\theta}_n-\bar{\theta}_n|+|\bar{\theta}_n-\theta_0|\leqslant|\hat{\theta}_n-\theta_0|+2|\bar{\theta}_n-\theta_0|$, 即得 θ_n^* 的强 (弱) 相合性.

(c) 上述结果还不能解释为 MLE 的强相合性, 除非可以验证 MLE 是似然方程的唯一解.

对于多维参数 $\theta=(\vartheta_1,\cdots,\vartheta_k)^\tau$ 的情形, 也有类似的结果, 但需要补加条件. 设总体 $X\sim dP_\theta(x)=f(x,\theta)d\mu(x)$, $X_1,\cdots,X_n$ 为从中抽取的 i.i.d. 样本, 参数空间 Θ 为 $\mathbf{R}^k$ 中的一个开集. 除上面的条件 (i), (ii) 之外, 我们还需要如下的条件

(iii) 对任一 $\theta^0\in\Theta$, 存在它的一个邻域 $U_{\theta^0}\subset\Theta$ 和一个正值函数 $H(x)$, 对 U_{θ^0} 中任意两点 θ,θ', 有

$$|\log f(x,\theta')-\log f(x,\theta)|\leqslant H(x)||\theta'-\theta||,\quad\text{而 }\mathrm{E}_{\theta^0}[H(X)]<\infty.\tag{5.18}$$

我们有如下定理.

定理 5.10 *假定上述条件成立, 则对任一 $\theta^0\in\Theta$, 在 P_{θ^0} 之下, 以概率 1 当 n 充分大时似然方程*

$$\frac{\partial\log L(X^n,\theta)}{\partial\theta}:=\sum_{i=1}^{n}\frac{\partial\log f(X_i,\theta)}{\partial\theta}=0\tag{5.19}$$

有解, 且所得解序列 $\{\hat{\theta}_n\}$ 是强相合的.

证明 取 $a>0$ 充分小, 使闭球 $B_a=\{\theta:||\theta-\theta^0||\leqslant a\}\subset U_{\theta^0}$, 将 B_a 的表面记为 S_a. 与一维参数情形的证明不同之处在于, 代替 (5.16), 此处需证: 对任一 $\theta^0\in\Theta$, 在 P_{θ^0} 之下, 以概率 1 当 n 充分大时有

$$\sup\Big\{\sum_{i=1}^n\log f(X_i,\theta):\theta\in S_a\Big\}<\sum_{i=1}^n\log f(X_i,\theta^0). \tag{5.20}$$

今证明如下:

1° 可以找到一串 $\theta_n\in S_a$ 收敛到某个 $\theta^*\in S_a$, 使

$$\mathrm{E}_{\theta^0}\log[f(X,\theta_n)/f(X,\theta^0)]\to\zeta:=\sup\{\mathrm{E}_{\theta^0}\log[f(X,\theta)/f(X,\theta^0)]:\theta\in S_a\}.$$

由于 $\log f(x,\theta_n)-\log f(x,\theta^*)\to 0$, 利用条件 (iii) 及控制收敛定理, 知

$$\mathrm{E}_{\theta^0}\log[f(X,\theta_n)/f(X,\theta^0)]\to\mathrm{E}_{\theta^0}\log[f(X,\theta^*)/f(X,\theta^0)]=\zeta,$$

由引理 1.28 即得 $\zeta=\mathrm{E}_{\theta^0}\log[f(X,\theta^*)/f(X,\theta^0)]<0$.

2° 记 $\delta=-\zeta>0$. 在 S_a 上可以找到一个有限集 V, 使对每个 $\theta\in S_a$, 存在 $v\in V$, $||\theta-v||<\delta/(8\mathrm{E}_{\theta^0}H(X))$. 由 1° 中的结果和强大数定律, 以概率 1 当 n 充分大时有 $\sum_{i=1}^n\log[f(X_i,v)/f(X_i,\theta^0)]<-n\delta/2$, 且由条件 (iii) 和强大数定律, 以概率 1 当 n 充分大时,

$$\begin{aligned}\Big|\sum_{i=1}^n\big[\log f(X_i,\theta)-\log f(X_i,v)\big]\Big|&\leqslant\sum_{i=1}^nH(X_i)||\theta-v||\\&<2n\mathrm{E}_{\theta^0}[H(X)]||\theta-v||<n\delta/4,\end{aligned}$$

因而 $\sum_{i=1}^n\log[f(X_i,\theta)/f(X_i,\theta^0)]<-n\delta/4$. 这就证明了 (5.20).

由 (5.20), 在 P_{θ^0} 之下, 以概率 1 当 n 充分大时, 似然函数在 B_a 内部有一局部极大, 它是似然方程的一个零点, 在这些零点中存在与 θ_0 最近的一个零点 $\hat{\theta}_n(a)$. 如果不考虑可测性问题, 与一维情形类似, 由此即可构造一列似然方程的强相合解. ■

例 5.8 设 $X_1,\cdots,X_n$ 为 X 的 i.i.d. 样本,

$$P_\theta(X=1)=1-P_\theta(X=0)=h(\theta)\equiv\begin{cases}\theta, & \theta\text{ 为有理数},\\ 1-\theta, & \theta\text{ 为无理数},\end{cases}$$

其中 $0<\theta<1$. 此例中不仅 θ 的极大似然估计不相合, 而且不存在相合估计.

前一结论留给读者证明, 现在我们来证明后一结论. 证明需要用到实变函数论中的一个事实, 即非空区间内的全体无理数不能表示为可列个闭集的并集, 也可应用泛函分析中的一个定理, 即一个完备距离空间不能表示为至多可列个无处稠密集的并集. 无处稠密集就是其闭包没有内点的集合. 一个集合如能表示为可列个闭集的并集, 则称为 F_σ 型集.

证明 用反证法. 假定 θ 有一列相合估计 $T_n=T(X_1,\cdots,X_n)$, 则不妨假定 T_n 有界, 由控制收敛定理, $\mathrm{E}_\theta(T_n)\to\theta, 0<\theta<1$. 记

$$h_n(\theta)=\sum_{x_i=0,1,i=1,\cdots,n}T_n(x_1,\cdots,x_n)\theta^{n\bar{x}}(1-\theta)^{n-n\bar{x}},$$

则 $h_n(\theta)\to h(\theta), 0<\theta<1$.

因 $h_n(\theta)$ 为 θ 的连续函数, 故对任一实数 a, $\{h_n \leqslant a\}\bigcap[1/2,3/4]$ 为闭集. 对任一 $a>0$ 有

$$\{h<a\}=\bigcup_{k=1}^{\infty}\bigcup_{m=1}^{\infty}\bigcap_{n\geqslant m}\{h_n\leqslant a-1/k\},$$

故 $\{h<a\}\bigcap[1/2,\ 3/4]$ 应为 F_σ 型集. 但若取 $a=1/2$, 则由

$$h(\theta):=\begin{cases}\theta, & \theta \text{ 为有理数}\\ 1-\theta, & \theta \text{ 为无理数},\end{cases}$$

$\{h<1/2\}\bigcap[1/2,3/4]$ 为区间 $[1/2,\ 3/4]$ 内的全体无理数的集合, 并由实变函数中的上述已知事实, 知它不是一个 F_σ 型集. 这个矛盾就表明参数 θ 不存在相合估计. ■

出发点不同的另一证明来自 Chen 和 Wu (1994). 该文研究了相合估计存在的必要条件, 并利用有关引理证明了上述结果.

一般说来, MLE 的一些深层次的渐近性质大多是在相合性的基础上发展起来的, 因此, **今后凡在讨论 MLE 的渐近性质时, 我们总是假定分布参数满足可识别性条件: 若 $\theta_1\neq\theta_2$, 则 $P_{\theta_1}\neq P_{\theta_2}$.**

2. 似然方程相合解的渐近有效性

下面我们讨论似然方程相合解的渐近有效性. 为此, 设总体 $X\sim dP_\theta(x)=f(x,\theta)d\mu(x)$, $\theta:=(\vartheta_1,\cdots,\vartheta_k)^\tau\in\Theta$, Θ 为 $\mathbf{R}^k$ 中的一个开集, $X_1,\cdots,X_n$ 为 i.i.d. 样本. 引进下面的规则性条件:

(i) 在 $\mathcal{X}\times\Theta$ 上, f 处处大于 0, 对 θ 有直到 2 阶的连续偏导数;

(ii) 对任一给定的 θ^0, 存在它的一个邻域 $U(\theta^0)$ 和可能依赖于该 θ^0 的函数 $g(x)>0$ 和 $G(x)>0$, 对该邻域内任一 θ 和任意的 $i,j\leqslant k$,

$$\Big|\frac{\partial f(x,\theta)}{\partial\vartheta_i}\Big|\leqslant g(x),\quad \Big|\frac{\partial^2 f(x,\theta)}{\partial\vartheta_i\partial\vartheta_j}\Big|\leqslant g(x),$$

$$\Big|\frac{\partial^2\log f(x,\theta)}{\partial\vartheta_i\partial\vartheta_j}-\frac{\partial^2\log f(x,\theta^0)}{\partial\vartheta_i\partial\vartheta_j}\Big|\leqslant G(x)||\theta-\theta^0||, \tag{5.21}$$

且

$$\int g(x)d\mu(x)<\infty,\quad \mathrm{E}_{\theta^0}[G(X)]<\infty;$$

(iii) Fisher 信息阵 $I(\theta):=\mathrm{E}_\theta S(X,\theta)S(X,\theta)^\tau$ 在 Θ 上存在且为正定, 此处, $S(x,\theta)\equiv\big(S_1(x,\theta),\cdots,S_k(x,\theta)\big)^\tau:=\partial\log f(x,\theta)/\partial\theta$ 称为计分函数.

我们有如下定理.

定理 5.11 *假定上面的规则性条件成立, 则对似然方程*

$$\frac{\partial\log L(X^n,\theta)}{\partial\theta}:=\sum_{i=1}^{n}\frac{\partial\log f(X_i,\theta)}{\partial\theta}=0$$

的任何一列相合解 $\{\hat\theta_n\}$, 当真参数为 θ^0 时, 在 P_{θ^0} 之下有

$$\sqrt{n}(\hat{\theta}_n-\theta^0)=I(\theta^0)^{-1}\frac{1}{\sqrt{n}}\sum_{i=1}^{n}S(X_i,\theta^0)+o_p(1), \tag{5.22}$$

因而

$$\sqrt{n}(\hat{\theta}_n-\theta^0)\rightsquigarrow N(0,I^{-1}(\theta^0)),\ n\to\infty. \tag{5.23}$$

证明 为简单计, 我们只对一维参数的情形给以证明, 多维情形的证明类似. 由条件 (i) 和 (ii),

$$\int\frac{\partial f(x,\theta)}{\partial\theta}d\mu(x)=\frac{\partial}{\partial\theta}\int f(x,\theta)d\mu(x)=\partial(1)/\partial\theta=0,\quad \int\frac{\partial^2 f(x,\theta)}{\partial\theta^2}d\mu(x)=0.$$

由此可得

$$\mathrm{E}_\theta[S(X,\theta)]=\mathrm{E}_\theta\left[\frac{\partial\log f(X,\theta)}{\partial\theta}\right]=0,$$

$$\begin{aligned}\mathrm{E}_\theta\left[\frac{\partial^2\log f(X,\theta)}{\partial\theta^2}\right]&=\int\left[\frac{1}{f(x,\theta)}\frac{\partial^2 f(x,\theta)}{\partial\theta^2}-\Big(\frac{1}{f(x,\theta)}\frac{\partial f(x,\theta)}{\partial\theta}\Big)^2\right]f(x,\theta)d\mu\\&=-\mathrm{E}_\theta[S(X,\theta)^2]=-I(\theta).\end{aligned}$$

由条件 (i) 和 (ii), 在 θ^0 的邻域 $U(\theta^0)$ 内有 Taylor 展开

$$S(x,\theta)=S(x,\theta^0)+(\theta-\theta^0)\frac{\partial^2\log f}{\partial\theta^2}(x,\theta^0)+\xi G(x)(\theta-\theta^0)^2,$$

此处 $|\xi|\leqslant 1$. 故有

$$(1/n)\partial\log L(X^n,\theta)/\partial\theta=A_n+B_n(\theta-\theta^0)+\xi_n C_n(\theta-\theta^0)^2,$$

此处, $|\xi_n|\leqslant 1$, 而

$$A_n=\frac{1}{n}\sum_{i=1}^{n}S(X_i,\theta^0),\ B_n=\frac{1}{n}\sum_{i=1}^{n}\frac{\partial^2\log f}{\partial\theta^2}(X_i,\theta^0),\ C_n=\frac{1}{n}\sum_{i=1}^{n}G(X_i).$$

由假定, $\{\hat{\theta}_n\}$ 为似然方程的一列相合解, 故

$$P_{\theta^0}\{\hat{\theta}_n\in U(\theta^0)\}\to 1.$$

而当 $\hat{\theta}_n\in U(\theta^0)$ 时, 有

$$0=\frac{1}{n}\frac{\partial\log L}{\partial\theta}(X^n,\hat{\theta}_n)=A_n+B_n(\hat{\theta}_n-\theta^0)+\xi_n^* C_n(\hat{\theta}_n-\theta^0)^2,$$

此处, $|\xi_n^*|\leqslant 1$. 由以上两式,

$$\sqrt{n}(\hat{\theta}_n-\theta^0)-\frac{-\sqrt{n}A_n}{B_n+\xi_n^* C_n(\hat{\theta}_n-\theta^0)}\xrightarrow{\text{pr.}}0.$$

由大数定律及 $\hat{\theta}_n\xrightarrow{\text{pr.}}\theta^0$, 有

$$B_n+\xi_n^* C_n(\hat{\theta}_n-\theta^0)\xrightarrow{\text{pr.}}\mathrm{E}_{\theta^0}\left[\frac{\partial^2\log f}{\partial\theta^2}(X,\theta^0)\right]=-I(\theta^0).$$

由此即得

$$\sqrt{n}(\hat{\theta}_n-\theta^0)=I(\theta^0)^{-1}\frac{1}{\sqrt{n}}\sum_{i=1}^{n}S(X_i,\theta^0)+o_p(1),$$

此即 (5.22). 再由 $\sqrt{n}A_n \rightsquigarrow N\big(0, I(\theta^0)\big)$ 及 Slutsky 定理, 即得

$$\sqrt{n}(\hat{\theta}_n - \theta^0) \rightsquigarrow N(0, I^{-1}(\theta^0)),\ n \to \infty.$$

这就是所要证明的. ■

当 X 的支撑与参数有关时, 定理的结论不成立. 单边或双边截断分布族就属于这种情形.

5.3 M-估计

理论研究和实践经验表明, 在数据分析特别是线性回归分析中常用的最小二乘方法, 当数据分布与正态或回归模型与线性回归有较大偏离时, 表现往往很不理想. 多年来, 统计学家提出了一些更加 "稳健" 的方法以供选择使用. M 方法即为其中之一, 并且可以说是最受重视、研究成果最为丰富的一种.

M 方法的研究发端于 Huber (1964) (1973). 这两篇论文分别研究了位置参数和线性回归的 M-估计及其渐近理论, 可以说是 M 方法的开山之作. 自此以后, 该领域一直受到统计学界的高度重视, 若干重要的研究成果相继问世. 有关专著和综述也不少, 例如可参看 Huber (1981), Rao 和 Toutenburg (1995), 陈希孺和赵林城 (1996), Jurečková 和 Sen (1996) 以及 Zhao (2000) (综述). 这些文献大多与回归分析有关. 本章主要研究的是通常的参数估计, 一般不涉及回归分析, 因而较多地参考了 van der Vaart (1998) 一书, 在此我们也将这本专著推荐给本书的读者.

5.3.1 M-估计的概念

设 $X_1, \cdots, X_n$ 为总体随机元 X 的 i.i.d. 样本, 我们关心的是与其分布关联的一个参数或泛函 θ, $\theta \in \Theta$. 需要指出的是, 与似然情形不同, 这里的参数可以是、也可以不是与分布一一对应的分布参数. 为估计参数 θ, 一种通用的方法是取一个合适的可测函数 $m_\theta : \mathcal{X} \mapsto \bar{\mathbf{R}} = [-\infty, \infty]$, 寻求目标函数

$$\theta \mapsto M_n(\theta) := \frac{1}{n}\sum_{i=1}^{n} m_\theta(X_i) \equiv \mathbb{P}_n m_\theta \tag{5.24}$$

在 Θ 上的最大值点 $\hat{\theta}_n = \hat{\theta}_n(X_1, \cdots, X_n) := \hat{\theta}_n(X^n)$, 以此作为真参数 θ^0 的一个估计, 称为**M-估计**. 此处, 我们可按通常方式配备 $\bar{\mathbf{R}}$ 上的 Borel σ-域 $\bar{\mathcal{B}}_1$.

例 5.9 (一维位置参数情形) 设

$$X = \theta + e, \quad e \sim dP(x) = f(x)d\mu(x), \quad -\infty < x, \theta < \infty,$$

其中 f 为一已知的密度函数, μ 为实轴上的一个 σ-有限测度, 参数 θ 称为位置参数. 设 $\{X_i = \theta + e_i, i = 1, \cdots, n\}$ 为 X 的 i.i.d. 观测值. 为估计真参数 θ^0, 目标函数通常可取为 $\mathbb{P}_n m(\cdot - \theta)$ 的形式, 即取 $m_\theta(x) = m(x - \theta)$.

在上述模型中设 e 的 p-分位数为 0, $0 < p < 1$, 则得分位数模型. 为估计真参数

θ^0, 定义

$$m(x)=x[p-I(x<0)]=\begin{cases}px, & \text{若 } x\geqslant 0,\\ (1-p)|x|, & \text{若 } x<0.\end{cases} \tag{5.25}$$

由分位数的定义, θ^0 为 X 的 p-分位数, 当且仅当函数 $\mathrm{E}[m(X-\theta)-m(X)]$ 在 θ^0 达到最小, 故可用 $\mathbb{P}_n m(\cdot-\theta)$ 的最小值点来估计, 其中 $m(x)$ 如 (5.25) 所示 (参看 1.4.1 节例 1.4). (我们附带指出, 此处 $\mathrm{E}|m(X-\theta)-m(X)|$ 总是有限的, 而 $\mathrm{E}|m(X-\theta)|<\infty$ 当且仅当 $\mathrm{E}|X|<\infty$.)

若取 $p=1/2$, 即知 X 的中位数 θ^0 可用 $\mathbb{P}_n|\cdot-\theta|$ 的最小值点估计, 称为**最小绝对偏差估计** (Least Absolute Deviations Estimates), 简记为 LAD 估计.

$\mathrm{E}(X)$ 的常用估计 $\bar{X}$ 即为 $\mathbb{P}_n(\cdot-\theta)^2$ 的最小值点, 这就是通常的**最小二乘估计** (Least Squares Estimates), 简记为 LS 估计. Huber 估计也是用于估计位置参数 θ 的一种M-估计, 与之相应的 m 函数为

$$m(x)=\begin{cases}-2kx-k^2, & \text{若 } x\leqslant -k,\\ x^2, & \text{若 } |x|<k,\\ 2kx-k^2, & \text{若 } x\geqslant k.\end{cases} \tag{5.26}$$

Huber 提出这一估计的初衷是为了限制极端值的影响, 选取合适的 k, 可以在估计的稳健性和渐近效率之间取得某种平衡.

由此可见, 常用的极大似然估计、最小二乘估计和最小绝对偏差估计等, 均可归入 M 估计的范畴. 事实上, M-估计构成了一个庞大的家族. 在估计未知参数时, 人们可以根据不同的使用目的和不同的分布特性选取合适的估计方法, 既可从中选取渐近效率高的估计, 也可选取适应性较强的稳健 (robust) 估计, 这给实际应用提供了较大的自由度. 我们还要再一次强调指出, M 方法在回归分析, 特别是在线性回归模型的统计研究中起着重要作用.

M-估计也可以定义为形如

$$\Psi_n(\theta):=\frac{1}{n}\sum_{i=1}^n\psi_\theta(X_i)\equiv\mathbb{P}_n\psi_\theta=0 \tag{5.27}$$

的估计方程的任何一个可测解, 此处 $x\mapsto\psi_\theta(x)=(\psi_{\theta,1}(x),\cdots,\psi_{\theta,k}(x))^\tau$ 为适当选取的可测的向量值函数. 这样定义的 M 估计, 在有些文献中, 也称之为 Z 估计. Z 者, 估计方程的零点也 (参看 van der Vaart (1998)).

对于这样一种定义, 在 $\Psi_n(\theta)$ 不是 θ 的连续函数的情形, 上述方程有时会没有解. 例如, 在例 5.9 中定义一维位置参数 θ 的 LAD 估计 $\hat{\theta}_n$ 时, 我们选取了凸函数 $\rho(u)=|u|$, 而将 $\hat{\theta}_n$ 定义为

$$\frac{1}{n}\sum_{i=1}^n\rho(X_i-\theta)\equiv\frac{1}{n}\sum_{i=1}^n|X_i-\theta|$$

在 Θ 上的最小值点. 函数 $\rho(u)=|u|$ 在 $u=0$ 处的左、右导数不等. 当用估计方程定义 θ 的 M-估计时, 文献中常取 $\psi(u)=\mathrm{sgn}(u)$, 即 (5.27) 中的 $\psi_\theta(x)=\mathrm{sgn}(x-\theta)$.

但此时估计方程

$$\Psi_n(\theta) := \frac{1}{n}\sum_{i=1}^{n}\operatorname{sgn}(X_i-\theta)=0$$

常常没有解. 因此有时要采取一些变通的办法. 例如像下面定理 5.13 中那样, 当 $n\to\infty$ 时, 允许估计序列 $\hat{\theta}_n$ 满足类似于 $|\Psi_n(\hat{\theta}_n)|=o_p(1)$ 的条件.

当然, 当采用估计函数的最大值点定义 M 估计时, 也会遇到类似的问题.

5.3.2 M-估计的相合性

由于 M 估计通常没有显式表达, 其确切分布难于求得, 因而对其渐近分布的研究就显得尤为重要.

本节首先讨论M-估计的相合性. 为此, 假定 P 为样本空间上的底分布, 需要估计的参数真值为 θ^0. 为使估计量 $\hat{\theta}_n$ 具有相合性, 并不需要 $\hat{\theta}_n$ 是目标函数 $M_n(\theta)$ 的最大值点, 只要其满足条件

$$M_n(\hat{\theta}_n)\geqslant \sup_{\theta} M_n(\theta)-o_p(1),$$

由此推得 $M_n(\hat{\theta}_n)\geqslant M_n(\theta^0)-o_p(1)$, 这后一条件配以其他条件, 已足以保证估计的相合性. 我们有如下定理.

定理 5.12 *设 M_n 为一随机函数. 假定存在一非随机函数 $M(\theta)$, 使得*

$$\sup_{\theta\in\Theta}|M_n(\theta)-M(\theta)|\xrightarrow{P}0,\quad n\to\infty, \tag{5.28}$$

且对每个 $\epsilon>0$,

$$\sup\{M(\theta):||\theta-\theta^0||\geqslant\epsilon\}<M(\theta^0), \tag{5.29}$$

则对满足 $M_n(\hat{\theta}_n)\geqslant M_n(\theta^0)-o_p(1)$ 的任一估计序列 $\hat{\theta}_n$, 必有 $\hat{\theta}_n\xrightarrow{P}\theta^0$.

证明 由 (5.28), $M_n(\theta^0)\xrightarrow{P}M(\theta^0)$, 再由 $\hat{\theta}_n$ 的性质, $M_n(\hat{\theta}_n)\geqslant M(\theta^0)-o_p(1)$, 因而,

$$\begin{aligned}M(\theta^0)-M(\hat{\theta}_n)&\leqslant M_n(\hat{\theta}_n)-M(\hat{\theta}_n)+o_p(1)\\&\leqslant\sup_{\theta}|M_n(\theta)-M(\theta)|+o_p(1)\xrightarrow{P}0.\end{aligned} \tag{5.30}$$

由 (5.29), 对每个 $\epsilon>0$, 存在一个 $\eta>0$, 只要 $||\theta-\theta^0||\geqslant\epsilon$ 即有 $M(\theta)<M(\theta^0)-\eta$. 故事件 $\{||\hat{\theta}_n-\theta^0||\geqslant\epsilon\}\subset\{M(\hat{\theta}_n)<M(\theta^0)-\eta\}$, 由 (5.30), 后一事件的概率收敛于 0. ■

需要指出的是, 上述表达式中某些量在有些情形可能是不可测的, 此时的有关概率陈述可以依据外概率来理解.

对于向量值的随机函数 $\theta\mapsto\Psi_n(\theta)$, 其零点即为函数 $\theta\mapsto-||\Psi_n(\theta)||$ 的最大值点. 因此我们立即得到如下定理.

定理 5.13 *设 Ψ_n 为一随机的向量值函数, $\Psi(\theta)$ 为 θ 的一个非随机的向量值函数, 使得对每个 $\epsilon>0$,*

$$\sup_{\theta\in\Theta}||\Psi_n(\theta)-\Psi(\theta)||\xrightarrow{P}0,\quad n\to\infty,$$

$$\inf\{||\Psi(\theta)|| : ||\theta-\theta^0|| \geqslant \epsilon\} > 0 = ||\Psi(\theta^0)||,$$

则对满足 $\Psi_n(\hat{\theta}_n) = o_p(1)$ 的任一估计序列 $\hat{\theta}_n$, 必有 $\hat{\theta}_n \xrightarrow{P} \theta^0$.

当采用 (5.24) 中的 $M_n(\theta) = \mathbb{P}_n m_\theta$, 或 (5.27) 中的 $\Psi_n(\theta) = \mathbb{P}_n \psi_\theta$ 时, 要求序列 $|M_n - M|$ 或 $|\Psi_n - \Psi|$ a.s. 一致收敛, 等价于要求函数类 $\{m_\theta : \theta \in \Theta\}$ 或 $\{\psi(\theta) : \theta \in \Theta\}$ 为 Glivenko-Cantelli 类. 正如附录例 A.5 中所示, 保证后者成立的一组简单的充分条件为: (i) 上述函数类的指标集 Θ 为一紧距离空间 (甚至可以不是欧氏的); (ii) 对每个 x, $\theta \mapsto m_\theta(x)$ 或 $\theta \mapsto \psi_\theta(x)$ 为连续映照; (iii)对所有的 θ, 函数 m_θ 或 ψ_θ 分别被一个 P-可积函数所控制. 此时, $M(\theta) = Pm_\theta, \Psi(\theta) = P\psi_\theta$.

关于目标函数收敛的一致性要求, 远比相合性所需要的强得多, 有时需要用其他一些假定来代替, 下面的引理就是这样的一个例子.

引理 5.14 设 Θ 为一实数区间, Ψ_n 和 Ψ 分别为 θ 的随机和非随机函数, 对每个 θ, $\Psi_n(\theta) \xrightarrow{P} \Psi(\theta), n \to \infty$. 且设对每个 $\epsilon > 0$, 有 $\Psi(\theta_0 - \epsilon) < 0 < \Psi(\theta_0 + \epsilon)$. 假定函数 $\theta \mapsto \Psi_n(\theta)$ 连续且恰好有一个零点 $\hat{\theta}_n$; 或者 Ψ_n 非减且 $\Psi_n(\hat{\theta}_n) = o_p(1)$, 则 $\hat{\theta}_n \xrightarrow{P} \theta_0$.

证明 函数 Ψ_n 连续且恰好有一个零点的情形, 有关相合性的证明留给读者作为练习. 今设 Ψ_n 非减且 $\Psi_n(\hat{\theta}_n) = o_p(1)$, 对任意给定的 $\epsilon > 0$, $\hat{\theta}_n \leqslant \theta_0 - \epsilon$ 蕴含 $\Psi_n(\hat{\theta}_n) \leqslant \Psi_n(\theta_0 - \epsilon)$, 故由此及 $\Psi(\theta_0 - \epsilon) < 0$, 有

$$\begin{aligned} P\{\hat{\theta}_n \leqslant \theta_0 - \epsilon\} &\leqslant P\{\Psi_n(\hat{\theta}_n) \leqslant \Psi(\theta_0 - \epsilon)/2\} + P\{\Psi_n(\theta_0 - \epsilon) \geqslant \Psi(\theta_0 - \epsilon)/2\} \\ &\leqslant P\{|\Psi_n(\hat{\theta}_n)| \geqslant |\Psi(\theta_0 - \epsilon)|/2\} \\ &\quad + P\{\Psi_n(\theta_0 - \epsilon) - \Psi(\theta_0 - \epsilon) \geqslant |\Psi(\theta_0 - \epsilon)|/2\}, \end{aligned}$$

由定理假定, 当 $n \to \infty$ 时上式右方趋于 0; 同样可得 $P\{\hat{\theta}_n \geqslant \theta_0 + \epsilon\} \to 0$. 这就证明了 $\hat{\theta}_n \xrightarrow{P} \theta_0$. ■

例 5.10 (中位数) 设总体 X 有唯一的中位数 $\mathrm{med}(X) = \theta_0$, 可以用样本中位数 $\hat{\theta}_n$ 估计它. $\hat{\theta}_n$ 为非减函数 $\theta \mapsto \Psi_n(\theta) = n^{-1}\sum_{i=1}^n \mathrm{sgn}(\theta - X_i)$ 的一个 ("近似") 零点, "近似" 一词可以由 $\Psi_n(\hat{\theta}_n) = o_p(1)$ 界定. 由大数定律, 对任一 θ,

$$\Psi_n(\theta) \xrightarrow{P} \Psi(\theta) = \mathrm{E}\ \mathrm{sgn}(\theta - X) = P(X < \theta) - P(X > \theta).$$

由假定, θ_0 为极限函数的唯一零点, 因而对每个 $\epsilon > 0$, 有 $\Psi(\theta_0 - \epsilon) < 0 < \Psi(\theta_0 + \epsilon)$. 故 $\hat{\theta}_n$ 的相合性很容易由引理 5.14 得出.

5.3.3 M-估计的收敛速度

本节将要在某些条件下, 导出 M-估计的收敛速度 $O_p(1/\sqrt{n})$, 这一结果不仅在推导极限分布时要用到, 而且还有其自身的意义. 为此, 设 $X_1, \cdots, X_n$ 为总体随机元 $X \sim P$ 的 i.i.d. 样本, $\mathbb{P}_n$ 为经验分布, 对某个距离空间 (Θ, d) 中的每个 θ, $x \mapsto m_\theta(x)$ 为一可测函数. 令 $\hat{\theta}_n$ 为目标函数 $\theta \mapsto M_n(\theta) := \mathbb{P}_n m_\theta$ 的一个 "近似的" 最大值点. 记 $G_n(\theta) := \mathbb{G}_n m_\theta \equiv \sqrt{n}(\mathbb{P}_n m_\theta - P m_\theta)$, 这就是通常所说的经验过程.

定理 5.15 假定

(1) 存在常数 $C>0$ 和 $\alpha>\beta>0$, 当 $\delta>0$ 充分小且 $d(\theta,\theta^0)<\delta$ 时

(1.1) $\quad P(m_{\theta^0}-m_\theta)\geqslant Cd(\theta,\theta^0)^\alpha$,

且对所有的 n, 有

(1.2) $\quad \mathrm{E}^*\sup_{d(\theta,\theta^0)<\delta}|\mathbb{G}_n(m_\theta-m_{\theta^0})|\leqslant C\delta^\beta$.

(2) 序列 $\hat{\theta}_n$ 满足 $\mathbb{P}_n m_{\hat{\theta}_n}\geqslant \mathbb{P}_n m_{\theta^0}-O_p\big(n^{-\alpha/(2\alpha-2\beta)}\big)$.

(3) 对任意给定的 $\epsilon>0$, 当 $n\to\infty$ 时, 有 $P^*\{d(\hat{\theta}_n,\theta^0)\geqslant\epsilon\}\to 0$.

则 $n^{1/(2\alpha-2\beta)}d(\hat{\theta}_n,\theta^0)=O_p^*(1)$.

定理中的假定 (1.2) 即为经验过程的极大不等式, 其证明有较大的难度. 定理中 $\alpha=2,\beta=1$ 的情形相当于通常的规则情形, $\{\sqrt{n}(\hat{\theta}_n-\theta^0)\}$ 依概率有界. 定理也适用于非标准情形, 例如, 当 $\alpha=2,\beta=1/2$ 时就会得到 $\{n^{1/3}(\hat{\theta}_n-\theta^0)\}$ 依概率有界的结果. 附录 A.4 中给出了一些极大不等式的例子, 可供在一定的应用范围内选用. 下面的推论 5.17, 就是为了建立 M-估计的渐近分布的目的, 应用经验过程极大不等式的一个例子.

为了使陈述和证明更加清晰, 我们将一个重要的特殊情形单独列为下面的定理并加以证明, 上述一般定理的证明与此并无区别. 定理 5.16 及其推论是建立 M-估计的渐近分布不可或缺的重要引理. 为简单计, 不妨设 Θ 即为欧氏空间的一个开子集.

定理 5.16 假定

(1) 存在常数 $C>0$, 当 $\delta>0$ 充分小且 $||\theta-\theta^0||<\delta$ 时

(1.1) $\quad P(m_{\theta^0}-m_\theta)\geqslant C||\theta-\theta^0||^2$,

且对所有的 n, 有

(1.2) $\quad \mathrm{E}^*\sup_{||\theta-\theta^0||<\delta}|\mathbb{G}_n(m_\theta-m_{\theta^0})|\leqslant C\delta$.

(2) 序列 $\hat{\theta}_n$ 满足 $\mathbb{P}_n m_{\hat{\theta}_n}\geqslant \mathbb{P}_n m_{\theta^0}-R_n$, 此处 R_n 为非负随机变量且 $nR_n=O_p(1)$;

(3) 对任意给定的 $\epsilon>0$, 当 $n\to\infty$ 时, 有 $P^*\{||\hat{\theta}_n-\theta^0||\geqslant\epsilon\}\to 0$.

则 $\sqrt{n}||\hat{\theta}_n-\theta^0||=O_P^*(1)$.

证明 对每个 n, $\Theta-\theta^0$ 可分解成形如 $S_{j,n}=\{\theta:2^{j-1}<\sqrt{n}||\theta-\theta^0||\leqslant 2^j\}$ 的球壳之并, 此处 j 跑遍整数集. 对给定的 $M>0$, 若 $\sqrt{n}||\hat{\theta}_n-\theta^0||>2^M$, 则对某个 $j>M$, $\hat{\theta}_n\in S_{j,n}$, 且 $\sup_{\theta\in S_{j,n}}\big(M_n(\theta)-M_n(\theta^0)\big)\geqslant -R_n$. 由此得出

$$\begin{aligned}P^*\{\sqrt{n}||\hat{\theta}_n-\theta^0||>2^M\}\leqslant&\sum_{j>M,2^j<\epsilon\sqrt{n}}P^*\{\sup_{\theta\in S_{j,n}}\big(M_n(\theta)-M_n(\theta^0)\big)\geqslant -K/n\}\\&+P^*\{2||\hat{\theta}_n-\theta^0||\geqslant\epsilon\}+P\{nR_n>K\}\\:=&\sum_{j>M,2^j<\epsilon\sqrt{n}}P^*\{A_{j,n}\}+I_n+J_n.\end{aligned}$$

由假定 (2), 对任意给定的 $\zeta>0$, 可取 K 充分大, 使得对所有的 n, $J_n<\zeta$; 而由假定 (3), 当 n 充分大时 $I_n<\zeta$. 选取 $\epsilon>0$ 充分小以保证对每个 $\delta\leqslant\epsilon$, 假定 (1) 成立,

则对上式和号中的每个 j,

$$\sup_{\theta\in S_{j,n}} \sqrt{n}\mathrm{E}[M_n(\theta^0)-M_n(\theta)] \geqslant C2^{2(j-1)}/\sqrt{n}.$$

当 $C2^{2(M-1)}>2K$ 时, 若 $A_{j,n}$ 成立, 则

$$-\inf_{\theta\in S_{j,n}} \sqrt{n}\big(M_n(\theta^0)-M_n(\theta)\big) \geqslant -K/\sqrt{n} \geqslant -C2^{2(j-1)}/(2\sqrt{n}).$$

因此,

$$\sum_{j>M,2^j<\epsilon\sqrt{n}} P^*\{A_{j,n}\} \leqslant \sum_{j>M,2^j<\epsilon\sqrt{n}} P^*\{\sup_{\theta\in S_{j,n}} |G_n(\theta)-G_n(\theta^0)| \geqslant C2^{2j-2}/(2\sqrt{n})\}.$$

由 Markov 不等式和条件 (1),

$$\sum_{j>M,2^j<\epsilon\sqrt{n}} P^*\{A_{j,n}\} \leqslant \sum_{j>M}(2^j/\sqrt{n})(2\sqrt{n}/2^{2j-2}) \leqslant \sum_{j>M} 1/2^{j-3}$$

取 M 充分大, 可使上式小于 ζ, 因而当 n 充分大时 $P^*\{\sqrt{n}||\hat{\theta}_n-\theta^0||>2^M\}<3\zeta$. ■

推论 5.17 设每个 θ 都在 $\mathbf{R}^k$ 的一个开子集内. 假定

(1) 对于每个 θ, $x\mapsto m_\theta(x)$ 为一可测函数, 且存在 θ^0 的一个开邻域和可测函数 $K(x), PK^2<\infty$, 使对该邻域中的所有 θ' 和 θ, 都有

$$|m_{\theta'}(x)-m_\theta(x)| \leqslant K(x)||\theta'-\theta||.$$

(2) θ^0 为函数 Pm_θ 的最大值点, 且该函数在 θ^0 附近有 2 阶的 Taylor 展开:

$$Pm_\theta = Pm_{\theta^0} + (1/2)(\theta-\theta^0)^\tau V(\theta^0)(\theta-\theta^0) + o(||\theta-\theta^0||^2).$$

其中 $V(\theta^0):=\partial^2 Pm_\theta/\partial\theta\partial\theta^\tau\big|_{\theta^0}<0$ 为负定.

(3) $\mathbb{P}_n m_{\hat{\theta}_n} \geqslant \sup_\theta \mathbb{P}_n m_\theta - O_p(1/n)$ 且 $\hat{\theta}_n \xrightarrow{P} \theta^0$.

则 $\sqrt{n}(\hat{\theta}_n-\theta^0)=O_p^*(1)$.

证明 由假定 (2), 定理 5.16 中的假定 (1.1) 成立. 为验证定理 5.16 的假定都成立, 只需证明其假定 (1.2) 成立. 为此, 我们将附录 A.4 中的推论 A.26 中的极大不等式用于函数类 $\mathcal{F}=\{m_\theta-m_{\theta^0}: ||\theta-\theta^0||<\delta\}$. 这个函数类有包络函数 $F=K\delta$, 故

$$\mathrm{E}^*\sup_{||\theta-\theta^0||<\delta}|\mathbb{G}_n(m_\theta-m_{\theta^0})| \leqslant C_0\int_0^{\delta||K||_{P,2}}\Big[\log N_{[]}\big(\epsilon,\mathcal{F},L_2(P)\big)\Big]^{1/2}d\epsilon.$$

式中 C_0 为一常数. 将附录 A.4 中的例 A.4 中关于类 $\mathcal{F}$ 的括号熵的上界代入, 即得

$$\begin{aligned}\mathrm{E}^*\sup_{||\theta-\theta^0||<\delta}|\mathbb{G}_n(m_\theta-m_{\theta^0})| &\leqslant C_1\int_0^{\delta||K||_{P,2}}\sqrt{\log(\delta/\epsilon)}d\epsilon\\ &= C_1\delta\int_{1/||K||_{P,2}}^{\infty} t^{-2}\sqrt{\log t}dt \leqslant C_2\delta.\end{aligned}$$

式中 C_1, C_2 为常数, 即 (1) 中第二个条件也成立.

由定理 5.16, 知推论中的结论成立. ■

5.3.4 M-估计的渐近正态性

借助于 Taylor 展开, 我们首先简单地展望一下M-估计渐近正态的可以预期的

结果. 参数 θ 的 M-估计可用最大化一个目标函数或解一个估计方程得到. 以第二种方法为例. 设 $X_1,\cdots,X_n$ 为取自分布 P 的 i.i.d. 样本, 为简单计, 假定 θ 是一维的. 设

$$\Psi_n(\theta):=\frac{1}{n}\sum_{i=1}^{n}\psi_\theta(X_i)\equiv\mathbb{P}_n\psi_\theta,\quad \Psi(\theta):=P\psi_\theta.$$

假定真参数 θ_0 为 Ψ 的一个零点, 估计 $\hat\theta_n$ 为 Ψ_n 的一个零点, 使 $\hat\theta_n\xrightarrow{P}\theta_0$, 则在 θ_0 的附近 $\Psi_n(\hat\theta_n)$ 可以有 Taylor 展开:

$$0=\Psi_n(\hat\theta_n)=\Psi_n(\theta_0)+\dot\Psi_n(\theta_0)(\hat\theta_n-\theta_0)+(1/2)\ddot\Psi_n(\tilde\theta)(\hat\theta_n-\theta_0)^2,$$

此处 $\tilde\theta$ 介于 $\hat\theta_n$ 和 θ_0 之间. 上式可改写为

$$\sqrt{n}(\hat\theta_n-\theta_0)=\frac{-\sqrt{n}\Psi_n(\theta_0)}{\dot\Psi_n(\theta_0)+(1/2)\ddot\Psi_n(\tilde\theta)(\hat\theta_n-\theta_0)}.\tag{5.31}$$

设 $0<P\psi_{\theta_0}^2<\infty$, 则分子 $-\sqrt{n}\Psi_n(\theta_0)=-n^{-1/2}\sum_{i=1}^n\psi_{\theta_0}(X_i)\rightsquigarrow N\big(0,P\psi_{\theta_0}^2\big)$, 此处用到了 $P\psi_{\theta_0}=\Psi(\theta_0)=0$. 至于分母, 由大数定律, 第一项 $\dot\Psi_n(\theta_0)\xrightarrow{P}P\dot\psi_{\theta_0}$, 第二项中因子 $\hat\theta_n-\theta_0=o_p(1)$, 在某些规则条件下, 可以期望另一因子 $\ddot\Psi_n(\tilde\theta)=O_p(1)$. 把这些合在一起, 由 Slutsky 定理, 即得

$$\sqrt{n}(\hat\theta_n-\theta_0)\rightsquigarrow N\big(0,\ P\psi_{\theta_0}^2[P\dot\psi_{\theta_0}]^{-2}\big).\tag{5.32}$$

上述启发式推导可以推广到 k-维参数的情形, 此时真参数 θ^0 为函数 $P\psi_\theta:\mathbf{R}^k\mapsto\mathbf{R}^k$ 的一个零点, $\hat\theta_n$ 为估计方程的一列相合解, 估计方程 $\Psi_n(\theta):=\mathbb{P}_n\psi_\theta=0$ 为 k 个方程组成的向量方程, 序列 $\dot\Psi_n(\theta^0):k\times k$, 依概率收敛于 $P\dot\psi(\theta^0):k\times k$. 与一维情形类似, 我们可以期望有

$$\sqrt{n}(\hat\theta_n-\theta^0)\rightsquigarrow N_k\Big(0,\big[P\dot\psi(\theta^0)\big]^{-1}P\psi_{\theta^0}\psi_{\theta^0}^\tau\big[P\dot\psi(\theta^0)^\tau\big]^{-1}\Big),\tag{5.33}$$

其中, $P\dot\psi(\theta^0)\equiv P\partial\psi/\partial\theta^\tau\big|_{\theta^0}$, 它的可逆性自然应列入假定之中.

但是推导中隐含的条件太强了! 例如, 对每个给定的 x, 函数 $\theta\mapsto\psi_\theta(x)$ 对参数 θ 要有直到 2 阶的连续偏导数. 在一些重要应用中, 例如在 $\psi_\theta(x)=\operatorname{sgn}(x-\theta)$ 的情形, 这个要求是无法达到的. 因此, 我们需要更加细致巧妙的分析, 用比此处弱得多的条件来建立M-估计的渐近正态性. 我们有以下的定理.

定理 5.18 假定

(i) Θ 为 $\mathbf{R}^k$ 的一个开子集, 对每个 $\theta\in\Theta$, $x\mapsto\psi_\theta(x)$ 为一可测的向量值函数;

(ii) 对以 θ^0 为中心的某个开球 $U_{\theta^0}\subset\Theta$, 存在一个可测函数 $K(x)$, $PK^2<\infty$, 对 U_{θ^0} 中任意的 θ,θ',

$$||\psi_{\theta'}(x)-\psi_\theta(x)||\leqslant K(x)||\theta'-\theta||;$$

(iii) θ^0 为函数 $\theta\mapsto P\psi_\theta:=\lambda(\theta)$ 的零点, 函数 λ 在 θ^0 处可微, 偏导数矩阵 $\partial\lambda/\partial\theta^\tau\big|_{\theta^0}:=\partial\lambda/\partial\theta^\tau(\theta^0)$ 非奇异, 且 $0<P||\psi_{\theta^0}||^2<\infty$;

(iv) $\mathbb{P}_n\psi_{\hat\theta_n}=o_p(n^{-1/2})$, 且 $\hat\theta_n\xrightarrow{P}\theta^0, n\to\infty$.

则当 $n \to \infty$ 时, 有

$$\sqrt{n}(\hat{\theta}_n - \theta^0) = -n^{-1/2}\Big[\frac{\partial \lambda}{\partial \theta^\tau}(\theta^0)\Big]^{-1} \sum_{i=1}^{n} \psi_{\theta^0}(X_i) + o_p(1), \tag{5.34}$$

因而有

$$\sqrt{n}(\hat{\theta}_n - \theta^0) \rightsquigarrow N_k\left(0, \Big[\frac{\partial \lambda}{\partial \theta^\tau}(\theta^0)\Big]^{-1} P\psi_{\theta^0}\psi_{\theta^0}^\tau \Big[\frac{\partial \lambda^\tau}{\partial \theta}(\theta^0)\Big]^{-1}\right).$$

证明 对任一可测函数 f, 记 $\sqrt{n}(\mathbb{P}_n - P)f := \mathbb{G}_n f$. 由条件 (ii), 以及附录 A.4 中的定理 A.24 和例 A.4, 函数类 $\mathcal{F} := \{\psi_\theta : \theta \in U_{\theta^0}\}$ 为一 P-Donsker 类, 故 $\{\mathbb{G}_n\psi_\theta : \theta \in U_{\theta^0}\}$ 在 $l^\infty(\mathcal{F})$ 中弱收敛于一 Gauss 过程 $\{\mathbb{G}_P\psi_\theta, \theta \in U_{\theta^0}\}$. 由条件 (ii), 当 $\hat{\theta}_n \in U_{\theta^0}$, $\int ||\psi_{\hat{\theta}_n}(x) - \psi_{\theta^0}(x)||^2 dP(x) \leqslant ||\hat{\theta}_n - \theta^0||^2 PK^2$, 故对任意给定的 $\epsilon > 0$, 由 $\hat{\theta}_n \xrightarrow{P} \theta^0$, 当 $n \to \infty$ 时有

$$\begin{aligned} & P\Big\{\int ||\psi_{\hat{\theta}_n}(x) - \psi_{\theta^0}(x)||^2 dP \geqslant \epsilon\Big\} \\ \leqslant\ & P\big\{\hat{\theta}_n \notin U_{\theta^0}\big\} + P\big\{||\hat{\theta}_n - \theta^0|| \geqslant (\epsilon/PK^2)^{1/2}\big\} \to 0. \end{aligned}$$

故由附录 A.4 中的引理 A.27, 有

$$\mathbb{G}_n(\psi_\theta - \psi_{\theta^0})\big|_{\hat{\theta}_n} \xrightarrow{P} 0. \tag{5.35}$$

考虑到 $\mathbb{P}_n\psi_{\hat{\theta}_n} = o_p(n^{-1/2})$ 和 $P\psi_{\theta^0} \equiv \lambda(\theta^0) = 0$, 即得

$$\sqrt{n}\partial\lambda/\partial\theta^\tau(\theta^0)(\hat{\theta}_n - \theta^0) + \sqrt{n}||\hat{\theta}_n - \theta^0||o_p(1) = \mathbb{G}_n\psi_{\theta^0} + o_p(1),$$

由此即得所要的结果. ■

因函数 $\theta \mapsto \mathrm{sgn}(x - \theta)$ 不满足 Lipschitz 条件, 故这种情形仍未能为定理条件所覆盖. 为了定理能够适用于估计中位数这一简单而又重要的情形, 我们注意到, 前一定理证明的关键在于要保证能够应用附录 A.4 中的引理 A.27 以得到 (5.35). 为此, 只要将上面的条件 (ii) 修改为以下的条件 (ii′). 与函数类 $\{I(x \leqslant \theta), \theta \in \mathbf{R}\}$ 一样, 函数类 $\{\mathrm{sgn}(x - \theta), \theta \in \mathbf{R}\}$ 满足条件 (ii′) (参看附录 A.4 中的例 A.3). 我们有如下定理.

定理 5.19 假定去掉上一定理的条件 (ii) 改用下面的条件 (ii′):

(ii′) 函数类 $\{x \mapsto \psi_\theta(x), \theta \in \Theta\}$ 为一 Donsker 类, 函数 $\theta \mapsto \psi_\theta$ 在 θ^0 处依概率连续, 且存在以 θ^0 为中心的一个开球 $U_{\theta^0} \subset \Theta$, 和一个可测函数 $K(x)$, $PK^2 < \infty$, 对此开球中的每个 θ 都有 $||\psi_\theta(x)|| \leqslant K(x)$.

且其他条件保持不变, 则上一定理的结论仍然成立.

函数 $\theta \mapsto \psi_\theta$ 在 θ^0 处**依概率连续**是指, 对于任给的 $\eta > 0$,

$$\lim_{\theta \to \theta^0} P\{||\psi_\theta(X) - \psi_{\theta^0}(X)|| \geqslant \eta\} = 0;$$

或者等价地, 对于任给的 $\eta > 0$, 和任意一列 $\theta_n \to \theta^0$,

$$\lim_{\theta_n \to \theta^0} P\{||\psi_{\theta_n}(X) - \psi_{\theta^0}(X)|| \geqslant \eta\} = 0.$$

证明 要证附录 A.4 中的引理 A.27 的条件全都满足, 只要证明

$$\int ||\psi_{\hat{\theta}_n}(x) - \psi_{\theta^0}(x)||^2 dP(x) \xrightarrow{P} 0.$$

为此, 记 $\omega=(X_1,X_2,\cdots)$. 由依概率连续的假定, 对于任意给定的 $\eta>0,\zeta>0,\exists\,\delta>0$, 只要 $||\theta-\theta^0||<\delta$ 即有

$$\rho(\theta,\theta^0):=P\big\{||\psi_\theta(X)-\psi_{\theta^0}(X)||\geqslant\eta\big\}<\zeta.$$

由 X 与 ω 独立, $P\{||\psi_{\hat\theta_n(\omega)}(X)-\psi_{\theta^0}(X)||\geqslant\eta|\omega\}=\rho\big(\hat\theta_n(\omega),\theta^0\big)$, 因而

$$\begin{aligned}&P\big\{||\hat\theta_n(\omega)-\theta^0||<\delta,||\psi_{\hat\theta_n(\omega)}(X)-\psi_{\theta^0}(X)||\geqslant\eta\big\}\\=&\int_{||\hat\theta_n(\omega)-\theta^0||<\delta}P\big\{||\psi_{\hat\theta_n(\omega)}(X)-\psi_{\theta^0}(X)||\geqslant\eta|\omega\big\}P(d\omega)<\zeta,\end{aligned}$$

故由 $\hat\theta_n\xrightarrow{P}\theta^0$,

$$\begin{aligned}&P\big\{(\omega,X):||\psi_{\hat\theta_n(\omega)}(X)-\psi_{\theta^0}(X)||\geqslant\eta\big\}\\\leqslant\,&P\big\{||\hat\theta_n-\theta^0||\geqslant\delta\big\}+P\big\{||\hat\theta_n(\omega)-\theta^0||<\delta,||\psi_{\hat\theta_n(\omega)}(X)-\psi_{\theta^0}(X)||\geqslant\eta\big\}\\\leqslant\,&P\big\{||\hat\theta_n-\theta^0||\geqslant\delta\big\}+\zeta<2\zeta\end{aligned}$$

对充分大的 n 成立, 因而

$$||\psi_{\hat\theta_n}(X)-\psi_{\theta^0}(X)||\xrightarrow{P}0,\quad n\to\infty.$$

由条件 (ii′) 及控制收敛定理即得

$$\mathrm{E}[I(\hat\theta_n\in U_{\theta^0})||\psi_{\hat\theta_n}(X)-\psi_{\theta^0}(X)||^2]\to0,$$

故对任意给定的 $\epsilon>0$,

$$\begin{aligned}&P\Big\{\int||\psi_{\hat\theta_n}(x)-\psi_{\theta^0}(x)||^2dP(x)\geqslant\epsilon\Big\}\\\leqslant\,&P\{\hat\theta_n\notin U_{\theta^0}\}+P\Big\{\hat\theta_n\in U_{\theta^0},I(\hat\theta_n\in U_{\theta^0})\int||\psi_{\hat\theta_n}(x)-\psi_{\theta^0}(x)||^2dP(x)>\epsilon\Big\}\\\leqslant\,&P\{\hat\theta_n\notin U_{\theta^0}\}+(1/\epsilon)\mathrm{E}[I(\hat\theta_n\in U_{\theta^0})||\psi_{\hat\theta_n}(X)-\psi_{\theta^0}(X)||^2]\to0,\end{aligned}$$

即 $\int||\psi_{\hat\theta_n}(x)-\psi_{\theta^0}(x)||^2dP(x)\xrightarrow{P}0$. 因而由附录 A.4 中的引理 A.27 即得 (5.35). ■

注 5.2 应用上述定理和引理 5.14, 可以非常简洁地建立样本分位数的 Bahadur 表示, 详见后面的引理 5.45.

下面我们讨论由最大化目标函数 $\mathbb{P}_nm_\theta$ 定义的 M-估计的渐近分布. 为此假定它为 θ^0 的相合估计, 其中 θ^0 为函数 $Pm_\theta:=u(\theta)$ 的最大值点. 如果该函数在 θ^0 两次连续可微, 则可有如下的 Taylor 展开:

$$u(\theta):=Pm_\theta=Pm_{\theta^0}+\frac{1}{2}(\theta-\theta^0)^\tau\frac{\partial^2u}{\partial\theta\partial\theta^\tau}(\theta^0)(\theta-\theta^0)+o(||\theta-\theta^0||^2).$$

下面的定理的适用范围涵盖了对中位数的估计.

定理 5.20 *设 Θ 是欧氏空间 $\mathbf{R}^k$ 的一个开子集. 假定*

(1) *对于每个 $\theta\in\Theta$, $x\mapsto m_\theta(x)$ 为一可测函数; 且依概率 P 或对几乎所有的 x, $\theta\mapsto m_\theta(x)$ 在 θ^0 处可微, 偏导数向量为 $\dot m_{\theta^0}(x):=\partial m_\theta(x)/\partial\theta\big|_{\theta^0}$.*

(2) *存在 θ^0 的一个开邻域和可测函数 $K(x),PK^2<\infty$, 使对该邻域中的所有 θ' 和 θ, 都有*

$$|m_{\theta'}(x)-m_\theta(x)|\leqslant K(x)||\theta'-\theta||.$$

(3) θ^0 为函数 Pm_θ 的一个最大值点, 且该函数在 θ^0 附近有上述 Taylor 展开, 其中 $V(\theta^0) := \partial^2 u/\partial\theta\partial\theta^\tau(\theta^0) < 0$ 为负定.

(4) $\mathbb{P}_n m_{\hat\theta_n} \geqslant \sup_\theta \mathbb{P}_n m_\theta - o_p(1/n)$ 且 $\hat\theta_n \xrightarrow{P} \theta^0$.

则有

$$\sqrt{n}(\hat\theta_n - \theta^0) = -V^{-1}(\theta^0) n^{-1/2} \sum_{i=1}^{n} \dot m_{\theta^0}(X_i) + o_p(1).$$

特别地,

$$\sqrt{n}(\hat\theta_n - \theta^0) \rightsquigarrow N_k\big(0, V^{-1}(\theta^0) P\dot m_{\theta^0} \dot m_{\theta^0}^\tau V^{-1}(\theta^0)\big).$$

证明 由条件 (1), (2) 和控制收敛定理, 有 $P\dot m_{\theta^0} = \partial P m_\theta/\partial\theta\big|_{\theta^0} = 0$, 其中第二个等式是因为 θ^0 为函数 Pm_θ 的极值点. 记 $\sqrt{n}(\bar\theta_n - \theta^0) := \bar h_n := -V^{-1}(\theta^0)\mathbb{G}_n \dot m_{\theta^0}$, 则 $\bar h_n = O_P^*(1)$.

在定理条件下, 附录 A.4 中的定理 A.29 和推论 5.17 成立. 由推论 5.17, $\hat h_n := \sqrt{n}(\hat\theta_n - \theta^0) = O_P^*(1)$. 故由附录 A.4 中的定理 A.29,

$$\mathbb{G}_n\Big(\sqrt{n}[m_{\theta^0 + \hat h_n/\sqrt{n}} - m_{\theta^0}] - \hat h_n^\tau \dot m_{\theta^0}\Big) \xrightarrow{P} 0,$$
$$\mathbb{G}_n\Big(\sqrt{n}[m_{\theta^0 + \bar h_n/\sqrt{n}} - m_{\theta^0}] - \bar h_n^\tau \dot m_{\theta^0}\Big) \xrightarrow{P} 0.$$

两式相减, 利用 Taylor 展开, 经化简即得

$$-(1/2)(\hat h_n - \bar h_n)^\tau V(\theta^0)(\hat h_n - \bar h_n) = n\mathbb{P}_n m_{\bar\theta_n} - n\mathbb{P}_n m_{\hat\theta_n} + o_p(1),$$

再由条件 (4),

$$0 \leqslant -(1/2)(\hat h_n - \bar h_n)^\tau V(\theta^0)(\hat h_n - \bar h_n) \leqslant o_p(1),$$

即得 $\sqrt{n}(\hat\theta_n - \theta^0) = \bar h_n + o_p(1) = -V^{-1}(\theta^0)\mathbb{G}_n \dot m_{\theta^0} + o_p(1)$, 因而定理得证. ■

5.3.5 再访 MLE

设 $X_1, \cdots, X_n$ 为总体随机元 X 的 i.i.d. 样本, $X \sim dP_\theta(x) = f_\theta(x)d\mu(x), \theta \in \Theta$. 如前所述, MLE $\hat\theta_n$ 定义为函数 $\theta \mapsto (1/n)\sum \log f_\theta(X_i) := \mathbb{P}_n \log f_\theta$ 在 Θ 上的最大值点. 前面已用经典方法讨论了 MLE 的渐近理论. 为了建立 MLE 的渐近正态性, 所加的条件与 $f_\theta(x)$ 对 θ 有直到 3 阶的连续偏导数相差无几. 下面我们将介绍 Le Cam 等人 1950 至 1970 年间的不同的研究视角和由此得出的精妙结果.

为此, 我们假定参数空间 Θ 为 $\mathbf{R}^k$ 的一个开子集, 且**分布参数满足可识别性条件: 若 $\theta_1 \neq \theta_2$, 则 $P_{\theta_1} \neq P_{\theta_2}$**.

为了叙述方便, 我们假定 $\theta^0 \in \Theta$ 为真参数, 且将真的分布 P_{θ^0} 简记为 P. 需要指出的是, 由于真参数是未知的, 有关定理的条件和结论都需要对任意的 $\theta^0 \in \Theta$ 成立. 为要建立极大似然估计的渐近正态性, 可以尝试将定理 5.20 应用于 $m_\theta = \log f_\theta$, 或将定理 5.18 和 5.19 应用于 $\psi_\theta = \dot l_\theta := \partial \log f_\theta/\partial\theta$, 即模型的计分函数. 下面的定理 5.22 可以视为定理 5.20 的另一种变化形式. 奇妙的是, 定理条件居然蕴含了关系式 $P_\theta \dot l_\theta = -I_\theta$ 以及函数 $\theta \mapsto P_{\theta^0} \log f_\theta$ 的两次可微性, 尽管 2 阶导数的存在性并

不是假定的一部分. 这种奇特效果也许是源于一个简单的事实, 即函数 $x \mapsto \sqrt{f_\theta(x)}$ 的 $L_2(\mu)$-模为 1. 因此, 均方可微的假定没有加之于函数 $\theta \mapsto \log f_\theta$, 而是加到了函数 $\theta \mapsto \sqrt{f_\theta}$ 上.

称统计模型 $\{P_\theta, \theta \in \Theta\}$ 在 $\theta^0 \in \Theta$ 是**均方可微的**, 如果存在一个可测的向量值函数 $\dot{l}_{\theta^0}$, 使得当 $\theta \to \theta^0$ 时, 有

$$\int \Big[\sqrt{f_\theta} - \sqrt{f_{\theta^0}} - (1/2)(\theta - \theta^0)^\tau \dot{l}_{\theta^0} \sqrt{f_{\theta^0}}\Big]^2 d\mu = o\big(||\theta - \theta^0||^2\big). \tag{5.36}$$

称统计模型 $\{P_\theta, \theta \in \Theta\}$ 是**均方可微的**, 如果对每个 $\theta^0 \in \Theta$, (5.36) 都成立.

注 5.3 统计模型的均方可微性在 MLE 的渐近最优理论中充当了重要的角色. 需要指出的是, 如果 $\partial f_\theta(x)/\partial\theta$ 对每个 x 都存在, 且 $f_\theta(x) \neq 0$, 则

$$\frac{\partial}{\partial\theta}\sqrt{f}_\theta = \frac{1}{2\sqrt{f}_\theta}\frac{\partial}{\partial\theta} f_\theta = \frac{1}{2}\Big(\frac{\partial}{\partial\theta}\log f_\theta\Big)\sqrt{f}_\theta.$$

因此, 积分中的 $\dot{l}_{\theta^0}$, 正是模型的计分函数 $S_\theta \equiv \partial \log f_\theta/\partial\theta$ 在 θ^0 的值, 且表达式 $I_\theta = P\dot{l}_\theta \dot{l}_\theta^\tau \equiv PS_\theta S_\theta^\tau$ 定义了 Fisher 信息矩阵, 这些都与符号所要表示的一致 (参看定理 5.11 的条件 (iii)). 但在 (5.36) 中并不要求 $\partial f_\theta/\partial\theta\big|_{\theta^0}$ 对每个 x 都存在.

(5.36) 可以改写为

$$\xi(\theta) := \int \Big[\frac{\sqrt{f_\theta} - \sqrt{f_{\theta^0}}}{||\theta - \theta^0||} - \frac{1}{2}\frac{(\theta - \theta^0)^\tau}{||\theta - \theta^0||}\dot{l}_{\theta^0}\sqrt{f_{\theta^0}}\Big]^2 d\mu \to 0, \ \text{当}\ \theta \to \theta^0.$$

任取 $\theta_n = \theta^0 + h_n/r_n$, r_n^2 为一列趋于无穷的整数, $h_n \to h \neq 0$. 则上述要求等价于: 对于任意的 $r_n \to \infty$, 任意的 "方向" 向量 h 和任意的 $h_n \to h$,

$$\int \Big[r_n\big(\sqrt{f_{\theta^0+h_n/r_n}} - \sqrt{f_{\theta^0}}\big) - (1/2)h_n^\tau \dot{l}_{\theta^0}\sqrt{f_{\theta^0}}\Big]^2 d\mu = o(1). \tag{5.37}$$

这一等价性是基于如下的事实: $\lim_{\theta\to\theta^0}\xi(\theta) = 0$ 当且仅当对任一序列 $\theta_n \to \theta^0$, 有 $\lim_{\theta_n\to\theta^0}\xi(\theta_n) = 0$; 另一方面, 一个数列收敛于 0, 当且仅当对于它的每个子列, 存在进一步的子列收敛于 0.

记 $P = P_{\theta^0}$. 令 $W_n := 2\big(\sqrt{f_{\theta^0+h_n/r_n}/f_\theta} - 1\big)$, 由 (5.37), r_nW_n 在 $L_2(P)$ 中收敛于 $h^\tau\dot{l}_{\theta^0}$, 故极限函数 $h^\tau\dot{l}_{\theta^0}$ a.s. P 唯一, 因 h 可以任意选取, 因而 $\dot{l}_{\theta^0}$ a.s. P 唯一.

在证明定理 5.22 之前, 我们需要证明以下的引理.

引理 5.21 *假定 Θ 为 $\mathbf{R}^k$ 的一个开子集, 且统计模型 $\{P_\theta, \theta \in \Theta\}$ 均方可微. 则在真分布 $P \equiv P_{\theta^0}$ 之下, $P\dot{l}_{\theta^0} = 0$, 且 Fisher 信息矩阵 $I_{\theta^0} = P\dot{l}_{\theta^0}\dot{l}_{\theta^0}^\tau$ 存在. 进一步, 对每个收敛序列 $h_n \to h$, 有*

$$\sum_{i=1}^{r_n^2}\log\frac{f_{\theta^0+h_n/r_n}}{f_{\theta^0}}(X_i) = \frac{1}{r_n}\sum_{i=1}^{r_n^2} h^\tau\dot{l}_{\theta^0}(X_i) - \frac{1}{2}h^\tau I_{\theta^0}h + o_p(1). \tag{5.38}$$

特别地, 当 $r_n^2 = n$ 时, 有

$$\sum_{i=1}^{n}\log\frac{f_{\theta^0+h_n/\sqrt{n}}}{f_{\theta^0}}(X_i) = \frac{1}{\sqrt{n}}\sum_{i=1}^{n} h^\tau\dot{l}_{\theta^0}(X_i) - \frac{1}{2}h^\tau I_{\theta^0}h + o_p(1).$$

证明 为行文简便计, 我们只对 $r_n^2 = n$ 的情形加以证明, 一般情形的证明完全一样. 简记 $f_n = f_{\theta^0+h_n/\sqrt{n}}, f = f_{\theta^0}, g = h^\tau \dot{l}_{\theta^0}$. 由 (5.37), 在 $L_2(\mu)$ 中有 $\sqrt{n}(\sqrt{f_n} - \sqrt{f}) \to (1/2)g\sqrt{f}$, 因而 $\sqrt{f_n} \to \sqrt{f}$. 由内积的连续性, 注意到 $\int f_n d\mu = \int f d\mu = 1$, 即得

$$Pg = \int (1/2)g\sqrt{f}2\sqrt{f}d\mu = \lim \int \sqrt{n}(\sqrt{f_n} - \sqrt{f})(\sqrt{f_n} + \sqrt{f})d\mu = 0.$$

因为 h 可以任意选取, 故 $P\dot{l}_{\theta^0} = 0$.

随机变量 $W_{ni} = 2[\sqrt{f_n/f}(X_i) - 1]$ 以 P-概率 1 有定义. 由 (5.37),

$$\mathrm{Var}\Big[\sum_{i=1}^n W_{ni} - (1/\sqrt{n})\sum_{i=1}^n g(X_i)\Big] \leqslant \mathrm{E}\big[\sqrt{n}W_{ni} - g(X_i)\big]^2 \to 0,$$

$$\mathrm{E}\Big[\sum_{i=1}^n W_{ni}\Big] = 2n\Big(\int \sqrt{f_n f}d\mu - 1\Big) = -n\int [\sqrt{f_n} - \sqrt{f}]^2 d\mu \to -(1/4)Pg^2.$$

由定义, $Pg^2 = \int g^2 dP = h^\tau I_{\theta^0} h$. 因为均值和方差都收敛于 0 的随机变量序列必依概率收敛于 0, 故由以上两式即得

$$\sum_{i=1}^n W_{ni} = \frac{1}{\sqrt{n}}\sum_{i=1}^n g(X_i) - \frac{1}{4}Pg^2 + o_p(1). \tag{5.39}$$

由 Taylor 展式, $\log(1+x) = x - (1/2)x^2 + x^2 R(2x)$, 此处 $\lim_{x\to 0} R(x) = 0$, 因而

$$\begin{aligned}\sum_{i=1}^n \log \frac{f_n}{f}(X_i) &= 2\sum_{i=1}^n \log\Big(1 + \frac{1}{2}W_{ni}\Big) \\ &= \sum_{i=1}^n W_{ni} - \frac{1}{4}\sum_{i=1}^n W_{ni}^2 + \frac{1}{2}\sum_{i=1}^n W_{ni}^2 R(W_{ni}).\end{aligned} \tag{5.40}$$

由于 $\mathrm{E}\big[\sqrt{n}W_{ni} - g(X_i)\big]^2 \to 0$, 故若记 $nW_{ni}^2 = g^2(X_i) + \zeta_{ni}$, 则 $\mathrm{E}|\zeta_{ni}| \to 0$. 因此 $(1/n)\sum_{i=1}^n \zeta_{ni} = o_p(1)$, 由此以及大数定律, 有

$$\sum_{i=1}^n W_{ni}^2 = \frac{1}{n}\sum_{i=1}^n g^2(X_i) + \frac{1}{n}\sum_{i=1}^n \zeta_{ni} \xrightarrow{P} Pg^2.$$

由 Markov 不等式,

$$\begin{aligned}nP\big(|W_{ni}| > \epsilon\sqrt{2}\big) &\leqslant nP\big(g^2(X_i) > n\epsilon^2\big) + nP\big(|\zeta_{ni}| > n\epsilon^2\big) \\ &\leqslant \epsilon^{-2}Pg^2 I(g^2 > n\epsilon^2) + \epsilon^{-2}\mathrm{E}|\zeta_{ni}| \to 0.\end{aligned}$$

故 $P\big(\max_{1\leqslant i\leqslant n}|W_{ni}| > \epsilon\big) \leqslant nP\big(|W_{ni}| > \epsilon\big) \to 0$, 这蕴含了 $\max_{1\leqslant i\leqslant n}|R(W_{ni})| \xrightarrow{P} 0$. 因此 (5.40) 中的最后一项依概率收敛于 0. 综合 (5.40) 以下各式, 即得

$$\sum_{i=1}^n \log \frac{f_n}{f}(X_i) = \sum_{i=1}^n W_{ni} - \frac{1}{4}Pg^2 + o_p(1),$$

再由 (5.39) 即得所要的结果. ■

定理 5.22 假定 Θ 为 $\mathbf{R}^k$ 的一个开子集, $\theta^0 \in \Theta$ 为真参数, 约定以下的概率运算都是在真分布 $P \equiv P_{\theta^0}$ 之下进行的. 假定统计模型 $\{P_\theta, \theta \in \Theta\}$ 在 θ^0 处为均方可微. 且设存在 θ^0 的一个邻域和一个可测函数 $K(x)$, $PK^2 < \infty$, 使得对于该邻域中的

任意两点 θ,θ', 都有

$$|\log f_\theta(x)-\log f_{\theta'}(x)|\leqslant K(x)||\theta-\theta'||.$$

进一步, 假定 Fisher 信息矩阵 I_{θ^0} 非奇异, $\mathbb{P}_n\log f_{\hat\theta_n}\geqslant\sup_\theta\mathbb{P}_n\log f_\theta-o_p(1/n)$ 且 $\hat\theta_n\xrightarrow{P}\theta^0$. 则

$$\sqrt{n}(\hat\theta_n-\theta^0)=I_{\theta^0}^{-1}\frac{1}{\sqrt{n}}\sum_{i=1}^n\dot l_{\theta^0}(X_i)+o_p(1).$$

特别地, 我们有

$$\sqrt{n}(\hat\theta_n-\theta^0)\rightsquigarrow N_k\big(0,I_{\theta^0}^{-1}\big).$$

证明 我们将证明, 这一定理是定理 5.20 的一个推论. 为此需要验证, 若取 $m_\theta=\log f_\theta$ 和 $V(\theta^0)=-I_{\theta^0}$, 则定理 5.20 的条件成立.

任取 $r_n\to\infty,h_n\to h$, r_n^2 为整数. 令

$$W_n=2\Big(\sqrt{f_{\theta^0+h_n/r_n}/f_{\theta^0}}-1\Big).$$

由均方可微的假定, r_nW_n 在 $L_2(P)$ 中收敛于 $h^\tau\dot l_{\theta^0}$, 因而依概率收敛, 故

$$r_n(\log f_{\theta^0+h_n/r_n}-\log f_{\theta^0})=2r_n\log(1+W_n/2)\xrightarrow{P}h^\tau\dot l_{\theta^0}.$$

再由加在函数 $\theta\mapsto\log f_\theta$ 上的 Lipschitz 条件, 应用控制收敛定理, 即知上述收敛可以加强为在 $L_2(P)$ 中的收敛, 即

$$P\Big[r_n\big(\log f_{\theta^0+h_n/r_n}-\log f_{\theta^0}\big)-h_n^\tau\dot l_{\theta^0}\Big]^2\to0.\tag{5.41}$$

这表明函数 $\theta\mapsto\log f_\theta$ 是依概率可微的, 这也正是定理 5.20 所要求的.

另一方面, 若记

$$Z_n:=\mathbb{G}_{r_n^2}\Big[r_n\big(\log f_{\theta^0+h_n/r_n}-\log f_{\theta^0}\big)-h_n^\tau\dot l_{\theta^0}\Big],$$

则 Z_n 均值为 0, 且由 (5.41), 其方差

$$\mathrm{Var}(Z_n)\leqslant P\Big[r_n\big(\log f_{\theta^0+h_n/r_n}-\log f_{\theta^0}\big)-h_n^\tau\dot l_{\theta^0}\Big]^2\to0,$$

因而 $Z_n\xrightarrow{P}0$. 注意到 $P\dot l_{\theta^0}=0$, 有

$$\sum_{i=1}^{r_n^2}\Big[\log f_{\theta^0+h_n/r_n}(X_i)-\log f_{\theta^0}(X_i)\Big]-r_n^{-1}\sum_{i=1}^{r_n^2}h_n^\tau\dot l_{\theta^0}(X_i)$$
$$-r_nP\Big[r_n(\log f_{\theta^0+h_n/r_n}-\log f_{\theta^0})\Big]\xrightarrow{P}0.$$

将此式与引理 5.21 中的 (5.38) 相结合, 即得

$$r_nP\Big[r_n(\log f_{\theta^0+h_n/r_n}-\log f_{\theta^0})\Big]\to-(1/2)h^\tau I_{\theta^0}h.$$

由于 r_n,h_n 和 h 都是任意的, 因而函数 $\theta\mapsto P\log f_\theta$ 两次可微, 具有 2 阶偏导数矩阵 $-I_{\theta^0}$, 或者至少有如下的 2 阶 Taylor 展式:

$$P\log f_\theta=P\log f_{\theta^0}-(1/2)(\theta-\theta^0)^\tau I_{\theta^0}(\theta-\theta^0)+o(||\theta-\theta^0||^2),\theta\to\theta^0.$$

这就是我们想要证明的. ■

注 5.4 如果定理条件中只有均方可微性和 Lipschitz 条件成立, 则由引理 5.21 和附录 A.4 中的定理 A.29, 在真分布 $P \equiv P_{\theta^0}$ 之下, $P\dot{l}_{\theta^0} = 0$, Fisher 信息矩阵 $I_{\theta^0} = P\dot{l}_{\theta^0}\dot{l}_{\theta^0}^{\tau}$ 存在, 且有上面的 2 阶 Taylor 展式. 由此及附录 A.4 中的定理 A.29, 对任一给定的常数 $c > 0$, 和依概率有界的任一随机序列 $\hat{h}_n$, 有

$$\sup_{||h||\leqslant c}\Big|\sum_{i=1}^{n}\log\frac{f_{\theta^0+h/\sqrt{n}}}{f_{\theta^0}}(X_i) - h^{\tau}\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\dot{l}_{\theta^0}(X_i) + \frac{1}{2}h^{\tau}I_{\theta^0}h\Big| \xrightarrow{P_{\theta^0}} 0,$$

$$\sum_{i=1}^{n}\log\frac{f_{\theta^0+\hat{h}_n/\sqrt{n}}}{f_{\theta^0}}(X_i) - \hat{h}_n^{\tau}\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\dot{l}_{\theta^0}(X_i) + \frac{1}{2}\hat{h}_n^{\tau}I_{\theta^0}\hat{h}_n \xrightarrow{P_{\theta^0}} 0.$$

这两个式子在后面常常会用到.

下面我们给出能使均方可微得以成立的一些简单条件, 通常需要假定对几乎每个 x 函数 $\theta \mapsto f_\theta(x)$ 的可微性, 再加上一个控制性条件. 这项工作自然离不开积分收敛定理. 下面的引理就是一个这样的定理.

引理 5.23 (Vitali) 设 g_n 和 g 为任意可测函数, g_n 关于测度 μ 几乎处处或依测度收敛于 g, 且对某个 $r \geqslant 1$, $\limsup\int|g_n|^r d\mu \leqslant \int|g|^r d\mu < \infty$, 则 $\int|g_n - g|^r d\mu \to 0$.

证明 由假定, $0 \leqslant 2^{r-1}(|g_n|^r + |g|^r) - |g_n - g|^r \to 2^r|g|^r$, a. e. μ. 由 Fatou 引理和定理假定,

$$\begin{aligned}\int 2^r|g|^r d\mu &\leqslant \liminf\int\big[2^{r-1}\big(|g_n|^r + |g|^r\big) - |g_n - g|^r\big]d\mu\\ &\leqslant 2^r\int|g|^r d\mu - \limsup\int|g_n - g|^r d\mu,\end{aligned}$$

故 $\limsup\int|g_n - g|^r d\mu \leqslant 0$. 引理由此得证. ■

定理 5.24 设 $X \sim dP_\theta(x) = f_\theta(x)d\mu(x), \theta \in \Theta$, Θ 为 $\mathbf{R}^k$ 的一个开子集. 假定对每个 x, 函数 $\theta \mapsto w_\theta(x) = \sqrt{f_\theta(x)}$ 连续可微, 矩阵函数 $I_\theta = \int\big(\dot{f}_\theta/f_\theta\big)\big(\dot{f}_\theta^{\tau}/f_\theta\big)f_\theta d\mu$ 有定义且在 Θ 上连续, 则函数 $\theta \mapsto \sqrt{f_\theta(x)}$ 均方可微, 即当 $h \to 0$ 时

$$\int\Big[\big(\sqrt{f_{\theta+h}} - \sqrt{f_\theta}\big) - (1/2)h^{\tau}\dot{l}_\theta\sqrt{f_\theta}\Big]^2 d\mu = o(||h||^2), \tag{5.42}$$

且 $\dot{l}_\theta = \dot{f}_\theta/f_\theta$, a.s.$P_\theta$.

证明 由复合函数的微分规则, 对每个 x, 函数 $\theta \mapsto f_\theta(x) = w_\theta^2(x)$ 是连续可微的, 其梯度 $\dot{f}_\theta = 2w_\theta\dot{w}_\theta$. 对给定的 x, 作为 θ 的函数 $w_\theta(x)$ 是非负的, 因而 w 的零点必为 $\dot{w}$ 的零点. 故在 P_θ 之下, $\dot{f}_\theta/f_\theta = 2\dot{w}_\theta/w_\theta$ 当其分母为 0 时可以任意地定义, 因而 a. s. 有定义. 由假定, 矩阵函数 $\theta \mapsto I_\theta = 4\int\dot{w}_\theta\dot{w}_\theta^{\tau}d\mu$ 是连续的.

由在均方可积假定下 $\dot{l}_\theta$ 的 a.s.P_θ 唯一性, 故要证 (5.42) 当 $\dot{l}_\theta = \dot{f}_\theta/f_\theta = 2\dot{w}_\theta/w_\theta$ 时成立, 只要证明, 对于任意的 $r_n \to \infty$, 任意的 "方向" 向量 h 和任意的 $h_n \to h$,

$$\int\Big[r_n\big(w_{\theta+h_n/r_n} - w_\theta\big) - h^{\tau}\dot{w}_\theta\Big]^2 d\mu \to 0.$$

令 $g_n(x) := r_n\big(w_{\theta+h_n/r_n} - w_\theta\big)$. 因为函数 $\theta \mapsto w_\theta(x)$ 是连续可微的, 故 $g_n(x) = \int_0^1 h^{\tau}\dot{w}_{\theta+uh_n/r_n}(x)du$, 由 Jensen 不等式, $g_n^2(x) \leqslant \int_0^1(h_n^{\tau}\dot{w}_{\theta+uh_n/r_n}(x))^2 du$. 于是,

$$\begin{aligned}\limsup \int g_n^2(x)d\mu(x) &\leqslant \int\int_0^1 (h_n^\tau \dot{w}_{\theta+uh_n/r_n}(x))^2 du d\mu(x)\\ &= \int_0^1 \int (h_n^\tau \dot{w}_{\theta+uh_n/r_n}(x))^2 d\mu(x) du\\ &= (1/4)h_n^\tau\Big(\int_0^1 I_{\theta+uh_n/r_n}du\Big)h_n,\end{aligned}$$

此处用到了 Fubini 定理和 I_θ 的定义. 由 $h_n \to h$ 和函数 $\theta \mapsto I_\theta$ 局部的一致连续性, 右边收敛于 $(1/4)h^\tau I_\theta h = \int (h^\tau \dot{w}_\theta(x))^2 d\mu(x)$. 记 $g(x) = h^\tau \dot{w}_\theta(x)$, 则

$$\limsup \int g_n^2(x)d\mu(x) \leqslant \int g^2(x)d\mu(x) < \infty.$$

因为函数 $\theta \mapsto w_\theta$ 是连续可微的, 故由 $h_n \to h$, 有 $g_n(x) \to g(x)$ 对每个 x 成立. 由引理 5.23,

$$\int \Big[r_n\big(w_{\theta+h_n/r_n}(x) - w_\theta(x)\big) - h^\tau \dot{w}_\theta(x)\Big]^2 d\mu(x) = \int \big(g_n(x) - g(x)\big)^2 d\mu(x) \to 0.$$

这就证明了所要的结果. ■

在 X 的支撑为有限集且与参数无关的场合, 定理的条件可以放宽. 所得结果在后面讨论拟合优度检验时有用. 为此, 假定 X 有分布

$$P_\theta(X = a_i) = \pi_i(\theta) > 0,\ i = 1, \cdots, k,\ \theta \in \Theta,$$

此处, $a_1, \cdots, a_k$ 各不相同, $\theta = (\theta_1, \cdots, \theta_r)^\tau$.

定理 5.25 设 Θ 为 $\mathbf{R}^r$ 的一个开子集, Θ中不同的 θ 对应于不同的分布 P_θ, 偏导数 $\partial\pi_i(\theta)/\partial\theta_j$ 在 Θ 上存在连续, $1 \leqslant i \leqslant k, 1 \leqslant j \leqslant r$, 且 Fisher 信息阵 $I(\theta) = \big[I_{i,j}(\theta)\big]$ 为 r-阶正定阵, $I_{i,j}(\theta) = \sum_{v=1}^k \big(1/\pi_v(\theta)\big)\big[\partial\pi_v(\theta)/\partial\theta_i\big]\big[\partial\pi_v(\theta)/\partial\theta_j\big]$. 设 $X_1, \cdots, X_n$ 为 X 的 i.i.d. 样本, 则对任一 $\theta \in \Theta$, 在 P_θ 之下, 似然方程至少有一相合解, 且似然方程的任一相合解 $\hat{\theta}_n$ 必满足 $\sqrt{n}(\hat{\theta}_n - \theta) \rightsquigarrow N_r\big(0, I^{-1}(\theta)\big)$.

证明 以下的概率计算都是在 $P = P_\theta$ 之下进行的. 以 ν_{ni} 记 $X_1, \cdots, X_n$ 中等于 a_i 的个数, 记 $\nu_n = (\nu_{n1}, \cdots, \nu_{nk})^\tau, \pi(\theta) = \big(\pi_1(\theta), \cdots, \pi_k(\theta)\big)^\tau$, 则 ν_n 服从多项分布 $M\big(n; \pi(\theta)\big)$. 设 $V \sim M\big(1; \pi(\theta)\big)$, 则其分布可表达为

$$P_\theta\big(V = (v_1, \cdots, v_k)\big) := f(v, \theta) \equiv \prod_{i=1}^k \pi_i(\theta)^{v_i},$$

其中 $v_1, \cdots, v_k$ 跑遍其和为 1 的非负整数, $v = (v_1, \cdots, v_k)^\tau$. 故 ν_n 可以表示为随机元 V 的 i.i.d 样本 $V_1, \cdots, V_n$ 之和, 故由多维中心极限定理,

$$\sqrt{n}\big(\nu_n/n - \pi(\theta)\big) \overset{d.}{=} n^{-1/2}\sum_{i=1}^n \big(V_i - \pi(\theta)\big) \rightsquigarrow N_k(0, \Lambda), \tag{5.43}$$

其中 k-阶方阵 $\Lambda = Q(\theta) - \pi(\theta)\pi^\tau(\theta)$, $Q(\theta) := \mathrm{diag}\big(\pi_1(\theta), \cdots, \pi_k(\theta)\big)$. 且由定理条件及 $f(v, \theta)$ 的表达式, 定理 5.10 的条件满足, 故由该定理, 似然方程

$$\sum_{i=1}^k \frac{\nu_{ni}}{\pi_i(\theta)}\frac{\partial\pi_i(\theta)}{\partial\theta_j} = 0, 1 \leqslant j \leqslant r, \quad \text{即 } \frac{\partial\pi(\theta)^\tau}{\partial\theta}Q(\theta)^{-1}\nu_n/n = 0$$

存在一列解 $\hat{\theta}_n \to \theta$, a.s., 且 Fisher 信息阵$I(\theta) = \big(\partial\pi(\theta)^\tau/\partial\theta\big)Q(\theta)^{-1}\partial\pi(\theta)/\partial\theta^\tau$, 确如定理条件所示.

记 $\mathbb{1}_k = (1, \cdots, 1)^\tau \in \mathbf{R}^k$. 将 $\hat{\theta}_n$ 代入似然方程, 注意到 $\big(\partial\pi(\theta)^\tau/\partial\theta\big)\mathbb{1}_k = 0$, 有

$$\frac{\partial\pi(\hat{\theta}_n)^\tau}{\partial\theta}Q(\hat{\theta}_n)^{-1}\sqrt{n}\big(\nu_n/n - \pi(\theta)\big) = \sqrt{n}\frac{\partial\pi(\hat{\theta}_n)^\tau}{\partial\theta}Q(\hat{\theta}_n)^{-1}\big(\pi(\hat{\theta}_n) - \pi(\theta)\big).$$

由 π 具有连续的偏导数以及 $\hat{\theta}_n \to \theta$ a.s.,

$$\pi(\hat{\theta}_n) - \pi(\theta) = \partial\pi(\theta)/\partial\theta^\tau(\hat{\theta}_n - \theta) + o_p\big(||\hat{\theta}_n - \theta||\big). \tag{5.44}$$

将该式代入上式, 经过化简即得

$$\begin{aligned}&\frac{\partial\pi(\theta)^\tau}{\partial\theta}Q(\theta)^{-1}\frac{\partial\pi(\theta)}{\partial\theta^\tau}\sqrt{n}\big(\hat{\theta}_n - \theta\big) + o_p\big(\sqrt{n}||\hat{\theta}_n - \theta||\big)\\ =\ &\frac{\partial\pi(\theta)^\tau}{\partial\theta}Q(\theta)^{-1}\sqrt{n}\big(\nu_n/n - \pi(\theta)\big) + o_p\big(\sqrt{n}||\hat{\theta}_n - \theta||\big),\end{aligned}$$

故 $\sqrt{n}||\hat{\theta}_n - \theta|| = O_p(1)$, 因而

$$\sqrt{n}\big(\hat{\theta}_n - \theta\big) = I(\theta)^{-1}\frac{\partial\pi(\theta)^\tau}{\partial\theta}Q(\theta)^{-1}\sqrt{n}\big(\nu_n/n - \pi(\theta)\big) + o_p(1), \tag{5.45}$$

注意到 (5.43) 中关于多项分布的中心极限定理, 以及

$$\frac{\partial\pi^\tau}{\partial\theta}Q(\theta)^{-1}\big(Q(\theta) - \pi(\theta)\pi(\theta)^\tau\big)Q(\theta)^{-1}\frac{\partial\pi}{\partial\theta^\tau} = \frac{\partial\pi^\tau}{\partial\theta}Q(\theta)^{-1}\frac{\partial\pi}{\partial\theta^\tau} = I(\theta),$$

即得 $\sqrt{n}\big(\hat{\theta}_n - \theta\big) \rightsquigarrow N_r\big(0, I^{-1}(\theta)\big)$. ■

5.4 契合性

设 P_n 和 Q_n 为可测空间 $(\mathcal{X}_n, \mathcal{A}_n)$ 上的一对概率测度. 我们按照 Le Cam 和 Yang (1990), 称 (P_n, Q_n) 为一个二元试验. 设 T_n 为基于该试验的一个随机向量. 假定我们有一个二元试验序列 $\{(P_n, Q_n)\}$. 在许多场合, 我们需要知道 T_n 在 P_n 和 Q_n 之下的极限分布. 但就难度而言, 又往往是前易后难. 一个重要的问题是, 是否有可能通过前者的极限分布简便地求得后者的极限分布. 在假设检验中就常常会遇到这样的问题. Le Cam (1960) 提出的契合性 (contiguity) 概念和有关结果, 打开了解决问题的门径. 有关结果被 Hájek 和 Šidák (1967) 冠以 Le Cam 的第一到第三引理的名称而定格. 本章将介绍其中的第一和第三引理. 我们在参考文献中还列举了 Le Cam (1986), Le Cam 和 Yang (1990) 及其中文本 (1993), Hájek (1962) 等有关文献. 有关理论将按 van der Vaart (1998) 的模式加以介绍. 为此, 我们先回顾一下概率论中关于两个测度之间涉及绝对连续性、奇异性以及 Lebesgue 分解等方面的有关内容, 并通过类比引向有关的概念.

今设 P 和 Q 为一可测空间 $(\mathcal{X}, \mathcal{A})$ 上的两个测度. 此处我们只限于考虑有关 σ-域上的 σ-有限测度. 正如我们在第一章介绍过的, 如果对任何可测集 A, $P(A) = 0$ 蕴含了 $Q(A) = 0$, 则称 Q 关于 P 绝对连续, 或称测度 Q 受控于 P, 记为 $Q \ll P$.

另一方面, 如果 $\mathcal{X}$ 可以表示为两个互斥集 $\mathcal{X}_P$ 和 $\mathcal{X}_Q$ 之并, 使 $P(\mathcal{X}_Q)=Q(\mathcal{X}_P)=0$, 即测度 P 和 Q 分别集中于 $\mathcal{X}_P$ 和 $\mathcal{X}_Q$ 上, 则称 P 和 Q 相互正交或相互奇异, 记为 $P\perp Q$.

对于一般的两个测度 P 和 Q, 绝不仅限于上述两种关系之一. 为了搞清两者之间的关系, 假定 p 和 q 分别为测度 P 和 Q 关于某个测度 μ 的密度. 考虑集合

$$\mathcal{X}_P=\{p>0\},\quad \mathcal{X}_Q=\{q>0\}.$$

易见, 测度 P 和 Q 分别集中于 $\mathcal{X}_P$ 和 $\mathcal{X}_Q$ 上. 测度 Q 关于测度 P 有如下的 Lebesgue 分解:

$$Q=Q^a+Q^\perp,\quad Q^a(A)=Q\big(A\cap\{p>0\}\big),\quad Q^\perp(A)=Q\big(A\cap\{p=0\}\big),$$

其中, $Q^a\ll P, Q^\perp\perp P$, 分别称为 Q 关于 P 的绝对连续部分和正交 (或奇异) 部分. 此时, 对每个可测集 A, 有

$$Q^a(A)=\int_A (q/p)dP.$$

函数 q/p 是 Q^a 关于 P 的密度.

在统计中, 似然比是比 Radon-Nikodym 密度更为通用的名称. Q 关于 P 的似然比通常定义为 q/p 或 ∞, 视 $p>0$ 或 $p=0$ & $q>0$ 而定, 记为 dQ/dP (而不是 dQ^a/dP). 作为似然比, 通常只对它在概率 P 之下的概率性质感兴趣, 它是 a.s. 唯一确定的, $dQ/dP=q/p$, a.s. P, 因而为一个随机变量.

需要指出的是, dQ/dP 以及 Lebesgue 分解是与控制测度 μ 的选取无关的. 我们不加证明地引述以下的引理 (参看 van der Vaart (1998)).

引理 5.26 *设 P 和 Q 为两个概率测度, 关于某个测度 μ 分别具有密度 p 和 q, 则对上面定义的测度 Q^a 和 $Q^\perp$, 有*

(i) $Q=Q^a+Q^\perp, Q^a\ll P, Q^\perp\perp P$;

(ii) *对每个可测集 A, 有* $Q^a(A)=\int_A(q/p)dP$;

(iii) $Q\ll P$ *当且仅当* $Q(p=0)=0$ *当且仅当* $\int(q/p)dP=1$.

需要着重指出的是, 等式 $\int f dQ=\int f\cdot[dQ/dP]dP$ 不一定成立. 要它对每个可积函数 f 都成立, 测度 Q 必须关于 P 绝对连续. **但对任意的 P, Q 和任意的非负可测函数 f, 总有**

$$\int f dQ\geqslant\int_{p>0} fq d\mu=\int_{p>0} f\frac{q}{p}pd\mu=\int_{p>0} f\frac{dQ}{dP}dP.$$

这一不等式将会经常使用. 此处严格的不等式有可能会成立!

考虑一对概率测度 P 和 Q, Q 关于 P 绝对连续, 则随机向量 $X:\mathcal{X}\mapsto\mathbf{R}^k$ 的 Q-分布律可用对子 $(X^\tau, dQ/dP)^\tau$ 的 P-分布律 $P^{X,V}$ 表示为

$$\mathrm{E}_Q[f(X)]=\mathrm{E}_P\left[f(X)\frac{dQ}{dP}\right]\quad\text{或}\quad Q(X\in B)=\mathrm{E}_P\left[I_B(X)\frac{dQ}{dP}\right]=\int_{B\times\mathbf{R}} v dP^{X,V}(x,v).$$

这两个公式的正确性本质上依赖于 Q 对于 P 的绝对连续性, 因为 Q 关于 P 的奇异部分是无法通过任何变量的 P-分布律恢复的.

上述讨论启示我们思考这样一个重要的问题, 即在何种条件下, 一列随机向量 $X_n : \mathcal{X}_n \mapsto \mathbf{R}^k$ 的 Q_n-极限分布律可以通过相应的 P_n-极限分布律表达出来. 通过上面的分析, 要达此目标, 序列 $\{Q_n\}$ 必须在某种意义下 "渐近受控于" 序列 $\{P_n\}$. 下面定义的契合性就正是这样的一个概念.

定义 5.1 设有一个二元试验序列 $\{(P_n, Q_n)\}$. 如果对任一满足 $P_n(A_n) \to 0$ 的可测集序列 $\{A_n\}$, 必有 $Q_n(A_n) \to 0$, 则称序列 $\{Q_n\}$ 契合于 $\{P_n\}$ ($\{Q_n\}$ is contiguous to $\{P_n\}$), 或者称序列 $\{Q_n\}$ 渐近受控于 $\{P_n\}$, 记为 $Q_n \lhd P_n$. 如果 $Q_n \lhd P_n$ 和 $P_n \lhd Q_n$ 同时成立, 则称序列 $\{P_n\}$ 和 $\{Q_n\}$ 相互契合 ($\{P_n\}$ and $\{Q_n\}$ are contiguous), 或者称序列 $\{P_n\}$ 和 $\{Q_n\}$ 渐近互控, 记为 $P_n \lhd\rhd Q_n$.

在勒康, 罗昭容 (1993) 一书中, 上面的两句话分别称为 "$\{Q_n\}$ 跟随 $\{P_n\}$ 同居" 和 "$\{P_n\}$ 和 $\{Q_n\}$ 同居".

首先, 我们要根据 P_n 和 Q_n 的似然比的渐近行为给出刻画契合性的两个特征. 似然比 dQ_n/dP_n 和 dP_n/dQ_n 是非负的, 且

$$\mathrm{E}_{P_n}(dQ_n/dP_n) \leqslant 1, \quad \mathrm{E}_{Q_n}(dP_n/dQ_n) \leqslant 1,$$

故似然比序列 dQ_n/dP_n 和 dP_n/dQ_n 分别在 P_n 和 Q_n 之下是胎紧的, 由 Prohorov 定理, 每个子列都有弱收敛的子列. 另一方面, 由引理 5.26 的 (iii), 对两个概率测度 P 和 Q, 以下三个命题等价:

$$Q \ll P, \quad Q(dP/dQ = 0) = 0, \quad \mathrm{E}_P(dQ/dP) = 1.$$

这一等价性的渐近版本就体现在下面的引理中.

引理 5.27 (Le Cam 第一引理) 设 $\{(P_n, Q_n)\}$ 为一个二元试验序列, 则以下命题等价:

(i) $Q_n \lhd P_n$, 即对任一满足 $P_n(A_n) \to 0$ 的可测集序列 $\{A_n\}$, 必有 $Q_n(A_n) \to 0$.

(ii) 如果沿着某个子列 $dP_n/dQ_n \overset{Q_n}{\rightsquigarrow} U$, 则 $P(U > 0) = 1$.

(iii) 如果沿着某个子列 $dQ_n/dP_n \overset{P_n}{\rightsquigarrow} V$, 则 $\mathrm{E}(V) = 1$.

(iv) 对任何统计量 $T_n : \mathcal{X}_n \mapsto \mathbf{R}^k$, $T_n \xrightarrow{P_n} 0$ 蕴含了 $T_n \xrightarrow{Q_n} 0$.

证明 (i) 和 (iv) 的等价性可直接由渐近受控的定义得到: 给定统计量 T_n, 考虑集合 $A_n = \{||T_n|| > \epsilon\}$; 给定集合 A_n, 考虑统计量 $T_n = I_{A_n}$.

(i) $\Rightarrow$ (ii). 为简化记号, 就用 $\{n\}$ 记沿着它 $dP_n/dQ_n \overset{Q_n}{\rightsquigarrow} U$ 的子序列. 对每个 n, 记 $g_n(\epsilon) = Q_n(dP_n/dQ_n < \epsilon) - P(U < \epsilon)$, 由弱收敛的一揽子引理, 对每个 $\epsilon > 0$, 有 $\liminf g_n(\epsilon) \geqslant 0$, 因而可取一串 $\epsilon_n \downarrow 0$ 足够慢, 以致仍有 $\liminf g_n(\epsilon_n) \geqslant 0$. 于是若记 $A_n := \{dP_n/dQ_n < \epsilon_n \ \& \ q_n > 0\}$, 则有

$$P(U = 0) = \lim P(U < \epsilon_n) \leqslant \liminf Q_n(dP_n/dQ_n < \epsilon_n) = \liminf Q_n(A_n).$$

另一方面,

$$P_n(A_n) = P_n\Big(\frac{dP_n}{dQ_n} < \epsilon_n \ \& \ q_n > 0\Big) = \int_{dP_n/dQ_n < \epsilon_n} \frac{dP_n}{dQ_n} dQ_n \leqslant \int \epsilon_n dQ_n \to 0.$$

故由 (i), $P(U = 0) \leqslant \lim Q_n(A_n) = 0$.

(iii) ⇒ (i). 设 $P_n(A_n) \to 0$, 则 $I_{\mathcal{X}_n - A_n} \xrightarrow{P_n} 1$. 由于两个边缘序列的胎紧蕴含联合序列的胎紧, 故由 Prohorov 定理, $\{n\}$ 的每个子列都有进一步的子列, 沿着它 $(dQ_n/dP_n, I_{\mathcal{X}_n - A_n}) \overset{P_n}{\rightsquigarrow} (V, 1)$, 对某个弱极限 V. 因函数 $(v,t) \mapsto vt$ 连续且在集合 $[0,\infty) \times \{0,1\}$ 上非负, 故由弱收敛的一揽子引理,

$$\liminf Q_n\big(\mathcal{X}_n - A_n\big) \geqslant \liminf \int I_{\mathcal{X}_n - A_n} \frac{dQ_n}{dP_n} dP_n \geqslant \mathrm{E}1 \cdot V.$$

由 (iii) 上式右方为 $\mathrm{E}(V) = 1$, 左方也为 1, 因而 $Q_n(A_n) \to 0$.

(ii) ⇒ (iii). 对每个 n, P_n 和 Q_n 都受控于概率测度 $\mu_n := (P_n + Q_n)/2$, 两者关于 μ_n 的密度之和为 2, 每个密度都取值于区间 $[0,2]$. 由 Prohorov 定理, 每个子列都有进一步的子列, 沿着它

$$dP_n/dQ_n \overset{Q_n}{\rightsquigarrow} U, \quad dQ_n/dP_n \overset{P_n}{\rightsquigarrow} V, \quad W_n := dP_n/d\mu_n \overset{\mu_n}{\rightsquigarrow} W,$$

其中 U, V 和 W 为极限随机变量. 每个 W_n 在 μ_n 之下的期望都为 1, 故由有界序列 W_n 的弱收敛, 我们有 $\mathrm{E}(W) = 1$.

对区间 $[0,2]$ 上任给的连续函数 f, 定义函数 g 如次: 当 $0 \leqslant w < 2$ 时, $g(w) = f\big(w/(2-w)\big)(2-w)$; 而 $g(2) = 0$. 则 g 有界连续. 因 $dP_n/dQ_n = W_n/(2 - W_n)$, $dQ_n/dP_n = 2 - W_n$, 故由一揽子引理,

$$\begin{aligned}\mathrm{E}[f(U)] \leftarrow \mathrm{E}_{Q_n}\Big[f\Big(\frac{dP_n}{dQ_n}\Big)\Big] &= \mathrm{E}_{\mu_n}\Big[f\Big(\frac{dP_n}{dQ_n}\Big)\frac{dQ_n}{d\mu_n}\Big] \\ &= \mathrm{E}_{\mu_n}[g(W_n)] \to \mathrm{E}\Big[f\Big(\frac{W}{2-W}\Big)(2-W)\Big],\end{aligned}$$

此处, 当 $W = 2$ 时右边被积项的取值理解为 $g(2) = 0$. 取区间 $[0,2]$ 上的一串有界连续函数 $f_m \leqslant 1, f_m \downarrow I_{\{0\}}$, 由控制收敛定理,

$$P(U = 0) = \mathrm{E}[I_{\{0\}}(U)] = \mathrm{E}[I_{\{0\}}\big(W/(2-W)\big)(2-W)] = 2P(W = 0).$$

通过类似的论证, 对区间 $[0,2]$ 上每个连续函数 f, $\mathrm{E}[f(V)] = \mathrm{E}[f\big((2-W)/W\big)W]$, 此处, 当 $W = 0$ 时右边被积项的取值理解为 0. 取 $[0,2]$ 上一串非负连续函数 $f_m(x) \uparrow x$, 则由单调收敛定理,

$$\mathrm{E}(V) = \mathrm{E}[\big((2-W)/W\big)W] = \mathrm{E}[(2-W)I_{W>0}] = 2P(W > 0) - 1.$$

由最后两式即得 $P(U = 0) + \mathrm{E}(V) = 1$. ■

例 5.11 (渐近对数正态) 设 $Z \sim N(\mu, \sigma^2), Y = \mathrm{e}^Z$, 则称 Y 服从对数正态分布, 记为 $Y \sim \exp\{N(\mu, \sigma^2)\}$. 设 P_n 和 Q_n 为任意可测空间上的概率测度,

$$dQ_n/dP_n \overset{P_n}{\to} \exp\{N(\mu, \sigma^2)\}.$$

则 $P_n \lhd Q_n$. 进一步, $P_n \lhd\rhd Q_n$ 当且仅当 $\mu = -\sigma^2/2$.

事实上, 因极限变量 (对数正态) 恒取正值, 故由 Le Cam 第一引理中的 (ii), $P_n \lhd Q_n$. 而由引理中的 (iii), $Q_n \lhd P_n$ 当且仅当 $\mathrm{E}\{\exp[N(\mu, \sigma^2)]\} = 1$, 当且仅当 $\mu = -\sigma^2/2$.

以上虽是一个特殊的例子, 但在渐近理论中却有着重要的作用. 这种均值和方差的关系看起来有点特殊, 然而我们将会看到, 这正是在统计程序的渐近优良性理

论中自然出现的.

下面的定理是 Le Cam 第三引理的一般形式, 有关结果完全类似于非渐近情形, 完美地回答了前面提出的由 P_n-极限分布律计算 Q_n-极限分布律的问题.

定理 5.28 设对每个 n, P_n 和 Q_n 为可测空间 $(\mathcal{X}_n, \mathcal{A}_n)$ 上的概率测度, $X_n : \mathcal{X}_n \mapsto \mathbf{R}^k$ 为随机向量. 假定 $Q_n \lhd P_n$, 且

$$(X_n^\tau, dQ_n/dP_n)^\tau \overset{P_n}{\rightsquigarrow} (X^\tau, V)^\tau,$$

则 $X_n \overset{Q_n}{\rightsquigarrow} L$, 其中概率测度 L 由

$$L(B) := \mathrm{E}\big[I_B(X)V\big] \quad 或 \quad \int g(x)dL(x) = \mathrm{E}\big[g(X)V\big]$$

定义. 由此可知 L 的特征函数 $\psi_L(t) := \int \mathrm{e}^{it^\tau x} dL(x) = \mathrm{E}(\mathrm{e}^{it^\tau X}V)$.

证明 由 $V \geqslant 0$ 易证 L 为测度. 由 $Q_n \lhd P_n$, 知 $\mathrm{E}(V) = 1$, 故 L 为概率测度. 由 L 的定义立得 $\int g dL = \mathrm{E}\big[g(X)V\big]$ 对每个简单函数 g, 进而对非负可测函数乃至每个可积函数都成立.

今设 g 为任一非负连续函数, 则 $\mathbf{R}^k \times [0,\infty)$ 上的函数 $(x,v) \mapsto g(x)v$ 也是非负连续的. 由一揽子引理,

$$\liminf \mathrm{E}_{Q_n}[g(X_n)] \geqslant \liminf \int g(X_n)\frac{dQ_n}{dP_n}dP_n \geqslant \mathrm{E}\big[g(X)V\big].$$

反向应用一揽子引理, 即得 $X_n \overset{Q_n}{\rightsquigarrow} L$. ■

引理 5.29 (Le Cam 第三引理) 假定

$$\Big(X_n^\tau, \log\frac{dQ_n}{dP_n}\Big)^\tau \overset{P_n}{\rightsquigarrow} (X^\tau, Z)^\tau \sim N_{k+1}\Big(\begin{pmatrix}\mu\\ -\sigma^2/2\end{pmatrix}, \begin{pmatrix}\Sigma & \gamma\\ \gamma^\tau & \sigma^2\end{pmatrix}\Big),$$

则

$$X_n \overset{Q_n}{\rightsquigarrow} N_k(\mu+\gamma, \Sigma).$$

证明 引理所述是定理 5.28 的一个特殊情形. 由定理假定和连续映照定理, 序列 $(X_n^\tau, dQ_n/dP_n)^\tau \overset{P_n}{\rightsquigarrow} (X^\tau, \mathrm{e}^Z)^\tau$. 因 $Z \sim N(-\sigma^2/2, \sigma^2)$, 故 $Q_n \lhd\rhd P_n$. 按照上一定理, $X_n \overset{Q_n}{\rightsquigarrow} L$, L 的特征函数 $\psi_L(t) = \mathrm{E}\big(\mathrm{e}^{it^\tau X}\mathrm{e}^Z\big)$, 它等于 $(X^\tau, Z)^\tau$ 的特征函数在 $(t^\tau, -i)^\tau$ 的值, 即

$$\begin{aligned}\psi_L(t) &= \exp\Big\{it^\tau\mu - \frac{1}{2}\sigma^2 - \frac{1}{2}(t^\tau, -i)\begin{pmatrix}\Sigma & \gamma\\ \gamma^\tau & \sigma^2\end{pmatrix}\begin{pmatrix}t\\ -i\end{pmatrix}\Big\}\\ &= \exp\Big\{it^\tau(\mu+\gamma) - \frac{1}{2}t^\tau\Sigma t\Big\}.\end{aligned}$$

这正好是 $N_k(\mu+\gamma, \Sigma)$ 的特征函数. ■

例 5.12 (局部对立假设下的极限分布) 设 $X_1, \cdots, X_n$ 为 X 的 i.i.d. 样本, $X \sim dP_\theta(x) = f_\theta(x)d\mu(x), \theta \in \Theta$. 假定 Θ 为 $\mathbf{R}^k$ 的一个开子集, 且统计模型 $\{P_\theta, \theta \in \Theta\}$ 为均方可微. 由引理 5.21, 若 h 为 $\mathbf{R}^k$ 中一个给定的非零向量, 则对每个 $\theta \in \Theta$,

$$\sum_{i=1}^{n}\log\frac{f_{\theta+h/\sqrt{n}}}{f_\theta}(X_i)=\frac{1}{\sqrt{n}}\sum_{i=1}^{n}h^\tau\dot{l}_\theta(X_i)-(1/2)h^\tau I_\theta h+o_{P_\theta}(1).$$

上式左方即为 log 似然比 $\log dQ_n/dP_n$, 其中 $Q_n=P^n_{\theta+h/\sqrt{n}}$, $P_n=P^n_\theta$, 于是有

$$\log\frac{dP^n_{\theta+h/\sqrt{n}}}{dP^n_\theta}\overset{\theta}{\rightsquigarrow}N\big(-h^\tau I_\theta h/2,h^\tau I_\theta h\big).$$

因此, 由例 5.11, 分布序列 $P^n_{\theta+h/\sqrt{n}}$ 和 P^n_θ 相互契合. 借助于 Le Cam 第三引理, 只要知道了一列统计量在 θ 之下的极限分布, 我们即可得到该序列在局部参数 $\theta+h/\sqrt{n}$ 之下的极限分布. 事实上, T_n 常常可以有如下的渐近表示:

$$\sqrt{n}(T_n-\mu_\theta)=\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\psi_\theta(X_i)+o_{P_\theta}(1).$$

假定 ψ_θ 均值为 0, 二阶矩有限, 则由中心极限定理, $n^{-1/2}\sum(\psi_\theta(X_i)^\tau,\dot{l}_\theta(X_i)^\tau)^\tau$ 在 θ 之下渐近多元正态. 由 Slutsky 定理,

$$\Big(\sqrt{n}(T_n-\mu_\theta)^\tau,\log\frac{dP^n_{\theta+h/\sqrt{n}}}{dP^n_\theta}\Big)^\tau\overset{\theta}{\rightsquigarrow}N\Big(\begin{pmatrix}0\\-h^\tau I_\theta h/2\end{pmatrix},\begin{bmatrix}P_\theta\psi_\theta\psi_\theta^\tau & P_\theta\psi_\theta\dot{l}_\theta^\tau h\\ h^\tau P_\theta\dot{l}_\theta\psi_\theta^\tau & h^\tau I_\theta h\end{bmatrix}\Big).$$

最后, 应用 Le Cam 第三引理即可得出 $\sqrt{n}(T_n-\mu_\theta)$ 在 $\theta+h/\sqrt{n}$ 下的极限分布.

5.5 大样本检验

把假设检验归结为优化问题, 可解的情况并不多. 究其原因, 是当对立假设是复合假设时, 我们面临的优化常常是针对多个目标的 "一致" 优化. 因此, 在一般的情形, 人们只有借助于直观的考虑, 设法构造一个或一些看上去合理的检验. 例如, 若某个统计量在原假设下倾向于取较小的值, 而在对立假设下倾向于取较大的值, 则当该统计量的取值超过某个事先确定的临界值时就否定原假设的检验, 可视为合理的检验. 这个临界值的确定, 取决于检验统计量在原假设之下的分布. 但在许多场合, 我们无法求得其确切分布, 这时就只有求出它的极限分布. 这种基于检验统计量的极限分布的检验称为大样本检验. 本节主要研究一些重要的大样本检验.

5.5.1 似然比检验

设样本 $X\sim dP_\theta(x)=f(x,\theta)d\mu,\theta\in\Theta$, Θ_0 和 $\Theta-\Theta_0$ 都是 Θ 的非空子集. 要检验假设

$$H_0:\theta\in\Theta_0\leftrightarrow H_1:\theta\in\Theta_1=\Theta-\Theta_0. \tag{5.46}$$

定义统计量

$$LR(x):=\sup_{\theta\in\Theta}f(x,\theta)/\sup_{\theta\in\Theta_0}f(x,\theta), \tag{5.47}$$

它称为上述检验问题的**似然比**. 基于似然比统计量构作的检验称为**似然比检验** (likelihood ratio tests).

现在考虑 i.i.d. 样本的常用情形. 设有样本空间 $(\mathcal{X},\mathcal{B})$ 上的随机元 X, $dP_\theta(x)=f(x,\theta)d\mu:=f_\theta(x)d\mu$, $\theta\in\Theta$, μ 为一个 σ-有限测度, $X_1,\cdots,X_n$ 为 X 的 i.i.d. 样本. 需要检验假设 (5.46). 简记 $X^n=(X_1,\cdots,X_n), L(x^n)=\prod_{i=1}^n f(x_i,\theta)$. 按照上述定义, 此时的似然比统计量为

$$LR(X^n):=\sup_{\theta\in\Theta}L(X^n,\theta)/\sup_{\theta\in\Theta_0}L(X^n,\theta). \tag{5.48}$$

假定存在 $\hat\theta_n$ 和 $\hat\theta_{n0}$, 使

$$\hat\theta_n=\arg\max\{\prod_{i=1}^n f(X_i,\theta):\theta\in\Theta\},$$
$$\hat\theta_{n0}=\arg\max\{\prod_{i=1}^n f(X_i,\theta):\theta\in\Theta_0\},$$

则

$$LR(X^n)=\prod_{i=1}^n f(X_i,\hat\theta_n)/\prod_{i=1}^n f(X_i,\hat\theta_{n0}). \tag{5.49}$$

但有时我们只能找到似然方程的相合解, 或上述极值问题的 "近似" 解 (就像定理 5.22 中加在 $\hat\theta_n$ 上的条件所示的那样), 此时我们仍可用有关的解 $\hat\theta_n$ 和 $\hat\theta_{n0}$ 代入上式中, 用来作为上述检验问题的检验统计量, 记为 $LR_*(X^n)$. 当然, 为了所得检验的优良性, 我们最好选用极值问题或似然方程的渐近有效的解.

似然比检验是在 Neyman 和 Pearson (1928) 中提出的, 可以看成 Fisher 极大似然估计在检验问题中的引申, 两者有着类似的直观依据. 这种方法要求分子分母的极值存在有限, 且为 x 的 $\mathcal{B}$-可测函数. 在样本空间和参数空间均为欧氏空间的开子集的常用情形, 一般可假定函数 $f(x,\theta)$ 当 θ 给定时对 x 可测, 而当 x 给定时对 θ 连续, 即可保证这种可测性. 在大样本检验中, 检验统计量的提出一般是基于直观的考虑, 而检验的临界值则需要借助于检验统计量在原假设之下的极限分布才能近似地确定. 似然比统计量的极限分布较早是由 Wilks (1938) 和 Chernoff (1954) 得到的.

似然比的极限分布

设 (5.46) 中的 Θ 为 $\mathbf{R}^k$ 的一个有内点的子集, Θ_0 的维数 $r<k$. 当 $r>0$ 时, Θ_0 为 r-维的意思是, 存在 $\mathbf{R}^r$ 中有内点的子集 A, 及定义于 A 取值于 $\mathbf{R}^k$ 的连续可微的函数 g, 使 $\theta=g(\phi)$ 在 A 和 Θ_0 之间构成一一对应. 此时, 对 $\theta\in\Theta_0$, X 的分布也可用参数 ϕ 表示如下:

$$f_\theta(x)d\mu\equiv f(x,\theta)d\mu=f(x,g(\phi))d\mu\equiv f^*(x,\phi)d\mu:=f^*_\phi(x)d\mu.$$

设 $k\times r$ 的矩阵函数 $D(\phi)=\partial g/\partial\phi^\tau$ 的秩在 A 上恒为 r.

分别用 Θ_0° 和 A° 记 Θ_0 和 A 的内部. 我们引进下面的两个条件.

(i) 假定统计模型 $\{P_\theta,\theta\in\Theta\}$ 在每个 $\theta\in\Theta_0^\circ$ 均方可微, 即对每个这样的 θ, 存在一个可测的向量值函数 $\dot l_\theta$, 当 $\zeta\in\Theta$, $\zeta\to\theta$ 时, 有

$$\int\Big[\sqrt{f_\zeta}-\sqrt{f_\theta}-(1/2)(\zeta-\theta)^\tau\dot l_\theta\sqrt{f_\theta}\Big]^2d\mu=o\big(||\zeta-\theta||^2\big).$$

(参看 (5.36)). 进一步, 假定 Fisher 信息矩阵 I_θ 在 Θ_0° 上非奇异.

(ii) 假定对任一 $\theta \in \Theta_0^\circ$, 存在 θ 在 $\mathbf{R}^k$ 中的一个邻域和一个可测函数 $K(x)$, $P_\theta K^2 < \infty$, 使得对于该邻域中的任意两点 θ', θ'', 都有

$$|\log f_{\theta'}(x) - \log f_{\theta''}(x)| \leqslant K(x)||\theta' - \theta''||.$$

今设条件 (i) 成立, 则对任一 $\phi \in A^\circ$, 当 $\psi \to \phi$ 时, 有

$$\int \Big[\sqrt{f_{g(\psi)}} - \sqrt{f_{g(\phi)}} - (1/2)\big(g(\psi) - g(\phi)\big)^\tau \dot{l}_{g(\phi)}\Big]^2 dP_{g(\phi)} = o\big(||\psi - \phi||^2\big).$$

记 $\dot{l}_\phi^* = D^\tau(\phi)\dot{l}_{g(\phi)}$, 则由 g 的可微性的有关假定, 有

$$\int \Big[\sqrt{f_\psi^*} - \sqrt{f_\phi^*} - (1/2)\big(\psi - \phi\big)^\tau \dot{l}_\phi^* \sqrt{f_\phi^*}\Big]^2 d\mu = o\big(||\psi - \varphi||^2\big).$$

即有

(i′) 模型 $\{f_\phi^*(x)d\mu,\ \phi \in A^\circ\}$ 也是均方可微的, 且其 Fisher 信息矩阵 $I_\phi^* = D^\tau(\phi) I_{g(\phi)} D(\phi)$ 满秩.

另一方面, 由 g 的连续可微性, 对任一 $\phi \in A^\circ$, $r \times r$ 矩阵函数 $D^\tau(\psi)D(\psi)$ 的最大特征根在 ϕ 附近有上界, 故由 (ii), 可得

(ii′) 对任一 $\phi \in A$, 存在 ϕ 的一个邻域和一个可测函数 $K(x)$, $P_{g(\phi)}K^2 < \infty$, 使得对于该邻域中的任意两点 ψ', ψ'', 都有

$$|\log f_{\psi'}^*(x) - \log f_{\psi''}^*(x)| \leqslant K(x)||\psi' - \psi''||.$$

定理 5.30 *假定上述条件 (i) 和 (ii) 成立, 参数真值 $\theta = g(\phi) \in \Theta_0^\circ$, $X_1, \cdots, X_n$ 为从真分布 $P \equiv P_\theta$ 中抽取的 i.i.d. 样本, 统计量 $\hat{\theta}_n$ 和 $\hat{\phi}_n$ 满足条件*

$$\sqrt{n}(\hat{\theta}_n - \theta) = I_\theta^{-1} Z_n + o_p(1), \quad Z_n := \frac{1}{\sqrt{n}} \sum_{i=1}^n \dot{l}_\theta(X_i), \tag{5.50}$$

$$\sqrt{n}(\hat{\phi}_n - \phi) = {I_\phi^*}^{-1} \frac{1}{\sqrt{n}} \sum_{i=1}^n \dot{l}_\phi^*(X_i) + o_p(1), \tag{5.51}$$

即它们分别为有关参数的渐近有效估计. 记

$$LR_*(X^n) = \prod_{i=1}^n f(X_i, \hat{\theta}_n) / \prod_{i=1}^n f^*(X_i, \hat{\phi}_n),$$

则当 $n \to \infty$ 时, $Y_n \equiv 2\log LR_*(X^n) \rightsquigarrow \chi^2_{k-r}$.

注 5.5 在 $r = 0$ 即 Θ_0 退化为一点 $\theta^0 \in \Theta^0$ 的情形, 定理中的检验统计量应为 $LR_*(X^n) = \prod_{i=1}^n f(X_i, \hat{\theta}_n) / \prod_{i=1}^n f(X_i, \theta^0)$. 若假定条件 (i) 和 (ii) 当 $\theta = \theta^0$ 时成立, 且存在 $\hat{\theta}_n$ 满足 (5.50), 则当真参数 $\theta = \theta^0$ 时定理结论仍然成立. 证明也可大大简化.

应当指出, 如果我们采用定理 5.22 中的真正的或 "近似" 的极大似然估计, 以及满足条件的似然方程的解, 来构造检验统计量 $LR_*(X^n)$, 都会得到相同的极限分布.

证明 为明确起见, 我们假定 $0 < r < k$. 条件 (i) 和 (ii) 蕴含了 (i′) 和 (ii′). 正如 5.3.5 节注 5.4 指出的, 由这些条件, 应用引理 5.21 和附录 A.4 中的定理 A.29, $P\dot{l}_\theta = 0$, 且对依概率有界的任意随机序列 $\hat{\Delta}_n : k \times 1$ 和 $\hat{\delta}_n : r \times 1$, 有

$$\sum_{i=1}^{n}\log\frac{f_{\theta+\hat{\Delta}_n/\sqrt{n}}}{f_\theta}(X_i)=\hat{\Delta}_n^\tau\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\dot{l}_\theta(X_i)-\frac{1}{2}\hat{\Delta}_n^\tau I_\theta\hat{\Delta}_n+o_p(1). \tag{5.52}$$

$$\sum_{i=1}^{n}\log\frac{f^*_{\phi+\hat{\delta}_n/\sqrt{n}}}{f^*_\phi}(X_i)=\hat{\delta}_n^\tau\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\dot{l}^*_\phi(X_i)-\frac{1}{2}\hat{\delta}_n^\tau I^*_\phi\hat{\delta}_n+o_p(1). \tag{5.53}$$

令 $\hat{\Delta}_n:=\sqrt{n}(\hat{\theta}_n-\theta)$ 和 $\hat{\delta}_n:=\sqrt{n}(\hat{\phi}_n-\phi)$, 由 (5.52) 和 (5.50),

$$2\sum_{i=1}^{n}\log\frac{f_{\hat{\theta}_n}}{f_\theta}(X_i)=||I_\theta^{-1/2}n^{-1/2}\sum_{i=1}^{n}\dot{l}_\theta(X_i)||^2+o_p(1)=Z_n^\tau I_\theta^{-1}Z_n+o_p(1).$$

同样, 由 (5.53) 和 (5.51), 并考虑到 $\theta=g(\phi)$, $\dot{l}^*_\phi=D^\tau(\phi)\dot{l}_\theta$, 和 $I^*_\phi=D^\tau(\phi)I_\theta D(\phi)$, 经过简单的计算即得

$$\begin{aligned}2\sum_{i=1}^{n}\log\frac{f^*_{\hat{\phi}_n}}{f_\theta}(X_i)&=||I^{*-1/2}_\phi n^{-1/2}\sum_{i=1}^{n}\dot{l}^*_\phi(X_i)||^2+o_p(1)\\&=\big(I_\theta^{-1/2}Z_n\big)^\tau BI_\theta^{-1/2}Z_n+o_p(1).\end{aligned}$$

此处,

$$B:=I_\theta^{1/2}D(\phi)\big(D^\tau(\phi)I_\theta D(\phi)\big)^{-1}D^\tau(\phi)I_\theta^{1/2} \tag{5.54}$$

是秩为 r 的 k-阶对称幂等阵, 故由上面的两个式子, 有

$$2\log LR_*(X^n)=\big(I_\theta^{-1/2}Z_n\big)^\tau(I_k-B)I_\theta^{-1/2}Z_n+o_p(1), \tag{5.55}$$

此处, I_k 为一 k-阶恒等阵. 因 $I_\theta^{-1/2}Z_n\rightsquigarrow N_k(0,I_k)$, 且 I_k-B 是秩为 $k-r$ 的 k-阶对称幂等阵, 故由 χ^2 分布与幂等阵的关系 (参看 1.4.3 节), 知 $2\log LR_*(X^n)\rightsquigarrow\chi^2_{k-r}$. ■

下面的推论给出了似然比检验在局部对立假设之下的极限分布, 可用于比较同一水平不同检验在局部对立假设之下的渐近功效.

推论 5.31 *在定理条件下, 若 h 为 $\mathbf{R}^k$ 中一个给定的非零向量, 则当局部参数为 $\theta_n=\theta+h/\sqrt{n}$ 时, 在 P_{θ_n} 之下,*

$$2\log LR_*(X^n)\rightsquigarrow\chi^2_{k-r,\delta},$$

其中, $\delta^2=h^\tau I_\theta^{1/2}(I_k-B)I_\theta^{1/2}h$, 对称幂等阵 B 如 (5.54) 所示.

证明 此处我们沿用例 5.12 中的路线. 统计模型 $\{P_\theta,\theta\in\Theta\}$ 为均方可微. 由引理 5.21, 在 P_θ 之下,

$$\sum_{i=1}^{n}\log\frac{f_{\theta+h/\sqrt{n}}}{f_\theta}(X_i)=\frac{1}{\sqrt{n}}\sum_{i=1}^{n}h^\tau\dot{l}_\theta(X_i)-(1/2)h^\tau I_\theta h+o_p(1).$$

故 $\log\big(dP^n_{\theta+h/\sqrt{n}}/dP^n_\theta\big)\overset{\theta}{\rightsquigarrow}N\big(-h^\tau I_\theta h/2,h^\tau I_\theta h\big)$, 因而 $Q_n:=P^n_{\theta+h/\sqrt{n}}$ 和 $P_n:=P^n_\theta$ 相互契合. 由 (5.50), 在 P_n 之下, 有

$$I_\theta^{-1/2}Z_n=I_\theta^{-1/2}n^{-1/2}\sum_{i=1}^{n}\dot{l}_\theta(X_i),$$

$$\Big((I_\theta^{-1/2}Z_n)^\tau,\log\frac{dP^n_{\theta+h/\sqrt{n}}}{dP^n_\theta}\Big)^\tau\overset{\theta}{\rightsquigarrow}N\Big(\begin{pmatrix}0\\-h^\tau I_\theta h/2\end{pmatrix},\begin{pmatrix}I_k & I_\theta^{1/2}h\\ h^\tau I_\theta^{1/2} & h^\tau I_\theta h\end{pmatrix}\Big).$$

由 Le Cam 第三引理, 在 Q_n 之下, $I_\theta^{-1/2}Z_n \rightsquigarrow N_k\big(I_\theta^{1/2}h, I_k\big)$. 对于任一列随机变量 T_n, 由 Q_n 与 P_n 之间的相互契合性, 在 Q_n 下 $T_n = o_p(1)$ 当且仅当在 P_n 下 $T_n = o_p(1)$, 因而可以统一记为 $T_n = o_p(1)$. 故由 (5.55),

$$2\log LR_*(X^n) = \big(I_\theta^{-1/2}Z_n\big)^\tau (I_k - B) I_\theta^{-1/2} Z_n + o_{P_n}(1),$$

故 在 Q_n 之下, $2\log LR_*(X^n) \rightsquigarrow \chi^2_{k-r,\delta}$, 其中, $\delta^2 = h^\tau I_\theta^{1/2}(I_k - B) I_\theta^{1/2} h$. ■

例 5.13 (多样本情形) 设样本 $X_{i1}, \cdots, X_{in_i}$ i.i.d., 服从 $N(\mu_i, \sigma_i^2)$, $1 \leqslant i \leqslant m$, 且全部样本独立. 要检验的假设为

$$H : \sigma_1^2 = \cdots = \sigma_m^2 \ \leftrightarrow\ K : \sigma_1^2, \cdots, \sigma_m^2 \text{ 不完全相同.}$$

记 $S_i^2 = \sum_{j=1}^{n_i}(X_{ij} - \bar{X}_i)^2/n_i$, $\bar{X}_i = \sum_{j=1}^{n_i} X_{ij}/n_i$, $S^2 = \sum_{i=1}^m n_i S_i^2/n$, $n = \sum_{i=1}^m n_i$. 不难算得似然比为 $LR(X^n) = S^n / \prod_{i=1}^m S_i^{n_i}$,

$$2\log LR(X^n) = 2n\log S - 2\sum_{i=1}^m n_i \log S_i.$$

用证明定理 5.30 的方法, 稍加修正, 即可证得, 当原假设成立且 $\min\{n_1, \cdots, n_m\} \to \infty$ 时, $2\log LR(X^n) \rightsquigarrow \chi^2_{k-r} = \chi^2_{m-1}$. 因此, 对于给定的渐近水平 $\alpha \in (0,1)$, 大样本似然比检验的否定域为 $2\log LR(X^n) > \chi^2_{m-1}(\alpha)$.

应当注意的是, 根据定理 5.30 所作的大样本似然比检验 ϕ_n 只是保证了

$$\lim_{n\to\infty} P_\theta\{\text{在检验 } \phi_n \text{ 之下 } H \text{ 被否定}\} = \alpha, \quad \theta \in \Theta_0,$$

它并不能保证当 n 固定时该检验有水平 α, 甚至根本找不到一个常数 $c = c_n$, 使以 $2\log LR(X^n) > c_n$ 为否定域的检验具有给定的水平 α. 在例 5.13 中, 这一情况不会发生, 因为在原假设 H 之下 $2\log LR(X^n)$ 的分布与 θ 无关.

当 n 固定时, $2\log LR(X^n)$ 的分布与其极限分布 χ^2_{k-r} 有差距, 其均值方差也会有差距. 于是就有文献提出对 $2\log LR(X^n)$ 加以修正, 一种方法是作一线性变换 $2a_n \log LR(X^n) + b_n$, 使其均值方差即为 χ^2_{k-r} 的均值方差 $k-r$ 和 $2(k-r)$, 希望这样可以改善 $2\log LR(X^n)$ 的分布与其极限分布的近似程度. Bartlett (1937) 曾就本例提出一种修正方法, 即通常的 **Bartlett 修正**. 详见陈希孺 (1981, 1997) 中的例 3.6.2.

当检验问题中涉及的 Θ_0 和 Θ 有相同维数时, 例如需要检验的假设是 $\mu \leqslant \mu_0 \Leftrightarrow \mu > \mu_0$ 时, 定理 5.30 没有相应的结论; 而当分布族没有共同支撑时, 定理结论不成立. 后者的一个典型的例子就是截断分布族. 有趣的是, 在某些例子中, 似然比统计量的极限分布仍为 χ^2 分布, 只是自由度与定理结论不一样.

5.5.2 拟合优度检验

设 $X_1, \cdots, X_n$ 为取自某个总体 X 的 i.i.d. 样本, 要检验假设

$$H : X_1, \cdots, X_n \text{ 的公共分布为 } F, \tag{5.56}$$

其中 F 为一完全已知的分布, 或者要检验假设

$$H : X_1, \cdots, X_n \text{ 的公共分布属于 } \mathcal{F}, \tag{5.57}$$

此处 $\mathcal{F}$ 为一个给定的分布族, 例如正态分布族. 关于分布的这种统计假设可以是源于某种理论或假说, 因而对它所作的检验即可视为对这种理论或假说的一种检验. 以遗传学中著名的 Mendel 豌豆杂交试验为例. 豌豆是严格自花传粉、闭花授粉的植物, 这种作物可以大量繁殖, 且便于人工授粉以严格进行杂交试验. 在他的试验中, 曾用纯种的红花植株和白花植株进行杂交, 得到的子一代植株全开红花. 再让其自花传粉, 发现在所得子二代植株中红花与白花植株大体为 "3:1" 的比例. 诸如此类的大量试验, 是遗传学说中 Mendel 分离定律和自由组合定律的实验依据. 因此对 "3:1" 假说的检验, 可视为对 Mendel 分离定律的一个检验. 又如在统计数据分析中, 常常假定所得数据来自正态总体. 这可以有某种理论依据, 例如中心极限定理, 也可能是仅凭经验. 为要消除疑虑, 可以将其付诸检验.

此类问题当然可以归入一般检验问题的框架之中, 但若界定为 "**拟合优度检验**" 时, 则自有其不同于显著性检验的特点. 其一, 它没有一个明确的对立假设, 或者说, 可能的对立假设是全方位的; 其二, 检验不取 "接受" 与 "否定" 两者择一的方式, 在构造检验统计量及解释检验结果时, 只能以 "数据与假设 H 拟合到何种程度" 为考虑问题的着眼点和归属. 这一界定与 "拟合优度检验" 相符合.

1. X 只取有限个值且 F 完全已知的情形

设 (5.56) 中的 F 有如下形式:

$$P_\theta(X=a_i)=\pi_i>0,\ i=1,\cdots,k,$$

此处, $\pi=(\pi_1,\cdots,\pi_k)^\tau$ 已知, $a_1,\cdots,a_k$ 各不相同. 为检验假设 (5.56), 以 ν_{ni} 记 $X_1,\cdots,X_n$ 中等于 a_i 的个数, $\nu_n=(\nu_{n1},\cdots,\nu_{nk})^\tau$, 构作统计量

$$K_n=\sum_{i=1}^k\frac{(\nu_{ni}-n\pi_i)^2}{n\pi_i}\equiv\sum_{i=1}^k\frac{\nu_{ni}^2}{n\pi_i}-n. \tag{5.58}$$

直观想法如下: 若 H 成立, 则依大数定律有 $\nu_{ni}/n\approx\pi_i$, $(\nu_{ni}-n\pi_i)^2$ 应倾向于取较小的值, 因而加权和 $\sum_{i=1}^k c_{ni}(\nu_{ni}-n\pi_i)^2$ 是一个适当的检验统计量, 当它取大值时可认为数据与分布的拟合差. 因 $(\nu_{ni}-n\pi_i)/\sqrt{n}\rightsquigarrow N\big(0,\pi_i(1-\pi_i)\big)$, 故要使 K_n 有非退化的极限分布, 应取 $c_{ni}=c_i/n>0$, 但唯有在 $c_i=1/\pi_i$ 时极限分布有最简单的形式 χ^2_{k-1}. (5.58) 中的 ν_{ni} 和 $n\pi_i$ 分别称为 a_i 的 **经验频数**和**理论频数**. 无论从检验统计量的构作, 还是从下面的极限分布的结果来看, a_i 的具体值都不起作用, 它只是各个数据所属类别的一种标识.

下面的定理见于 K. Pearson (1900), 是数理统计学最重要的基础性结果之一. K. Pearson 则是现代统计的奠基人之一.

定理 5.32 (Pearson) 当假设 H 成立时, $K_n\rightsquigarrow\chi^2_{k-1},\ \ n\to\infty$.

证明 记 $\xi_{ni}=(\nu_{ni}-n\pi_i)/\sqrt{n\pi_i}$, $\xi_n=(\xi_{n1},\cdots,\xi_{nk})^\tau$, $\nu_n=(\nu_{n1}\cdots,\nu_{nk})^\tau$, $\pi=(\pi_1,\cdots,\pi_k)^\tau$, $Q=\mathrm{diag}(\pi_1,\cdots,\pi_k)$, $\psi=(\sqrt{\pi_1},\cdots,\sqrt{\pi_k})^\tau$, 则由 (5.43), 有

$$(\nu_n-n\pi)/\sqrt{n}\rightsquigarrow N_k(0,Q-\pi\pi^\tau),$$

$$\xi_n = Q^{-1/2}(\nu_n - n\pi)/\sqrt{n} \rightsquigarrow N_k(0, I_k - \psi\psi^\tau).$$

因 $\psi : k \times 1$ 为单位向量, 故存在矩阵 $A : k \times (k-1)$ 使 $(\psi\ A)$ 为 k 阶正交阵. 记 $\zeta_n = A^\tau \xi_n$, 由上述极限表达式、正交阵的性质, 以及 $\psi^\tau \xi_n = 0$, 有

$$\zeta_n \rightsquigarrow N_{k-1}(0, I_{k-1}) \quad \text{且 } \xi_n = \Big(\psi\ A\Big)\begin{pmatrix}\psi^\tau \\ A^\tau\end{pmatrix}\xi_n = \Big(\psi\ A\Big)\begin{pmatrix}0 \\ \zeta_n\end{pmatrix} = A\zeta_n,$$

故由 $A^\tau A = I_{k-1}$ 知

$$K_n = \xi_n^\tau \xi_n = \zeta_n^\tau A^\tau A \zeta_n = \zeta_n^\tau \zeta_n \rightsquigarrow \chi^2_{k-1}. \qquad \blacksquare$$

根据这个定理, 即可对 (5.56) 中的假设 H 作出如下的大样本检验: 对于给定的水平 α, 当 $K_n > \chi^2_{k-1}(\alpha)$ 时否定 H, 此检验有渐近水平 α. 另一种方法则是根据具体数据算出 K_n 的值 y, 再根据极限分布算出 $p(y) = P(\chi^2_{k-1} > y)$, 这种 p-值可以解释为当 H 成立而 n 很大时, 得到比 y 更大的偏差的近似概率. 如果这个概率较大, 则观测到的偏差 y 并不稀罕, 可以认为数据与理论的拟合是好的; 反之, 当这个概率很小时, 如果我们不接受 "发生了一个小概率事件", 则只好认为数据与理论拟合得不好. 基于这个分析, 可以把这个 p-值作为所观测到的数据与理论分布 F 拟合程度的一种度量, 称为 "**拟合优度**". 这个度量排除了 n 和 π_i 的影响, 是一个比较满意的选择. 当然, $p(y)$ 的计算与我们选用的检验统计量有关, 因而选用不同的统计方法会得出不同的结果. 但不依赖于所选用的方法的拟合优度是无法定义的.

需要提醒的一点是, 如果出现 "过度" 的拟合, 即算出的 p-值太大或 K_n 值太小, 则不能排除试验数据有人为干预的嫌疑.

在统计文献中, K_n 常被称为 (**Pearson**) χ^2**-统计量**, 以 K_n 大值为否定域的检验, 则称为 χ^2**-或 Pearson** χ^2**-拟合优度检验**.

2. X **只取有限个值而** F **带参数的情形**

仍设 X 只取 $a_1, \cdots, a_k$ 等 k 个不同的值. 原假设 H 是: 存在 $\theta \in \Theta$, X 有如下形式的分布

$$P_\theta(X = a_i) = \pi_i(\theta) > 0,\ 1 \leqslant i \leqslant k,$$

此处 $\theta = (\theta_1, \cdots, \theta_r)^\tau, 1 \leqslant r \leqslant k-2$, Θ 为 $\mathbf{R}^r$ 的一个开子集. 设 $X_1, \cdots, X_n$ 为 X 的 i.i.d. 样本, 以 ν_{ni} 记样本中等于 a_i 的 X_j 的个数, $\nu_n = (\nu_{n1}, \cdots, \nu_{nk})^\tau$. 为检验 H, 利用样本对 θ 作一估计 $\hat{\theta}_n$, 构作检验统计量

$$K_n = \sum_{i=1}^k \frac{\big(\nu_{ni} - n\pi_i(\hat{\theta}_n)\big)^2}{n\pi_i(\hat{\theta}_n)} \equiv \sum_{i=1}^k \frac{\nu_{ni}^2}{n\pi_i(\hat{\theta}_n)} - n. \tag{5.59}$$

定理 5.33 (Fisher) 设 Θ 为 $\mathbf{R}^r$ 的一个开子集, $\pi(\theta) = \big(\pi_1(\theta), \cdots, \pi_k(\theta)\big)^\tau$ 满足定理 5.25 的条件. 假定 $X \sim P_\theta, \theta \in \Theta$, 且 $\hat{\theta}_n$ 取为 θ 的似然方程的相合解. 则当 $n \to \infty$ 时, (5.59) 定义的 K_n 依分布收敛于 χ^2_{k-1-r}.

证明 记 $\hat{\pi}_i = \pi_i(\hat{\theta}_n)$, $\pi_i = \pi_i(\theta)$, $\hat{\pi} = (\hat{\pi}_1, \cdots, \hat{\pi}_k)^\tau$, $\pi = (\pi_1, \cdots, \pi_k)^\tau$, $Q = \mathrm{diag}(\pi_1, \cdots, \pi_k), B = Q^{-1/2}\partial\pi/\partial\theta^\tau : k \times r$, $\psi = (\sqrt{\pi_1}, \cdots, \sqrt{\pi_k})^\tau$, $\xi_{ni} = (\nu_{ni} -$

$n\pi_i)/\sqrt{n\pi_i}$, $\xi_n=(\xi_{n1},\cdots,\xi_{nk})^\tau$, 则 $B^\tau B=I(\theta)$, $\psi^\tau B=0$, 且

$$\xi_n \rightsquigarrow N_k(0, I_k-\psi\psi^\tau).$$

由 (5.44) 和 (5.45), 以及 K_n 的表达式, 在真分布 P_θ 之下有

$$\begin{aligned}\eta_n &:= \sqrt{n}Q^{-1/2}(\hat{\pi}-\pi)=\sqrt{n}Q^{-1/2}\partial\pi/\partial\theta^\tau(\hat{\theta}_n-\theta)+o_p(1)\\ &=\sqrt{n}B(\hat{\theta}_n-\theta)+o_p(1)=BI(\theta)^{-1}B^\tau\xi_n+o_p(1),\\ K_n &= ||\xi_n-\eta_n||^2+o_p(1)=||(I_k-BI(\theta)^{-1}B^\tau)\xi_n||^2+o_p(1).\end{aligned}$$

因 $B^\tau B=I(\theta)$, 故 $BI(\theta)^{-1}B^\tau$ 为对称幂等阵, 因而

$$K_n=\xi_n^\tau(I_k-BI(\theta)^{-1}B^\tau)\xi_n+o_p(1).$$

因 $\psi: k\times 1$ 为单位向量, 故存在矩阵 $A: k\times(k-1)$ 使 $(\psi\ A)$ 为 k 阶正交阵. 记 $\zeta_n=A^\tau\xi_n$, 由上述极限表达式, 正交阵的性质, 以及 $\psi^\tau\xi_n=0$, 与前一定理的证明一样, 有 $\xi_n=A\zeta_n$, 且

$$\zeta_n \rightsquigarrow N_{k-1}(0, I_{k-1}), \tag{5.60}$$

$$K_n=\zeta_n^\tau\big(A^\tau A-A^\tau BI(\theta)^{-1}B^\tau A\big)\zeta_n=\zeta_n^\tau\big(I_{k-1}-M\big)\zeta_n, \tag{5.61}$$

其中 $M=A^\tau BI(\theta)^{-1}B^\tau A$. 由 $AA^\tau=I_k-\psi\psi^\tau$, $\psi^\tau B=0$ 及 $B^\tau B=I(\theta)$, 有 $M^2=M$, 即 M 为对称幂等阵, 且由 $I(\theta)^{-1}B^\tau AA^\tau B=I(\theta)^{-1}B^\tau(I_k-\psi\psi^\tau)B=I(\theta)^{-1}B^\tau B=I_r$, 即得

$$\mathrm{tr}(M)=\mathrm{tr}\big(I(\theta)^{-1}B^\tau AA^\tau B\big)=\mathrm{tr}(I_r)=r.$$

由 (5.60), (5.61) 以及 $\mathrm{rank}(I_{k-1}-M)=k-1-r$, 知 $K_n \rightsquigarrow \chi^2_{k-1-r}$. ■

例 5.14 据遗传学理论, 人类血型由一基因位点上的三个等位基因 A,B,O 所决定, 它们产生六个基因型 OO,AO,AA,BO,BB,AB 和四个表现型: OO—血型 O; AO,AA—血型 A; BO,BB—血型 B; AB—血型 AB. 设有一个大的人类群体, 符合 Harday-Weinberg 平衡定律, 因而其基因频率及基因型频率保持稳定. 设群体中 A,B,O 三个基因的频率分别为 p,q 和 $s=1-p-q$, 则四种血型频率分别为

$$P(O)=s^2,\ P(A)=p^2+2ps,\ P(B)=q^2+2qs,\ P(AB)=2pq.$$

这相当于定理 5.33 中 $k=4, r=2$ 的情况. 设从该群体中随机抽取 n 个人, 其中血型为 O,A,B,AB 型的分别有 n_O,n_A,n_B,n_{AB} 个, 则 p,q,s 的 MLE $\hat{p}_n,\hat{q}_n,\hat{s}_n$ 为似然函数

$$\log L(p,q,s)=2n_o\log s+n_A\log(p^2+2ps)+n_B\log(q^2+2qs)+n_{AB}\log(2pq)$$

在约束条件 $p,q,s\geqslant 0, p+q+s=1$ 下的最大值点. 用迭代法得出 $\hat{p}_n,\hat{q}_n,\hat{s}_n$ 后, 即可按 (5.59) 算得 K_n 的值 y, 以及拟合优度 $P(\chi^2_1>y)$. 曾有人用不同地区和种族的人群数据做过这个检验, 与上述遗传学理论一般都有良好的拟合.

3. X 可取无穷个值的一般情形

一般地, 设 X 服从一个完全已知的分布 F, 或者其分布函数可表示为 $F_\theta,\theta\in\Theta$, Θ 为 $\mathbf{R}^r$ 的一个开子集. 处理方法是将 X 取值的区间 $(a_0,a_k]$ 分割为形如 $(a_{i-1},a_i]$

的 $k \geqslant r+2$ 个区间之并, $-\infty \leqslant a_0 < a_1 < \cdots < a_{k-1} < a_k \leqslant \infty$, 第 i 段的概率记为 $\pi_i(\theta) = F_\theta(a_i) - F_\theta(a_{i-1})$. 这样就将问题归结为已经处理过的情形.

Feller (1957) 第 6 章第 7 节列举了用 χ^2 方法拟合 Poisson 分布的若干例子, 我们选取了下面的这个例子.

例 5.15 (Rutherford 的 α 质点散射试验) 设某放射源放射出 α 质点, 以 X 记给定时间 t 内到达指定区域的质点数. 在试验中观察了 $n = 2608$ 次, 时间区间 $t = 7.5$ s, 表 5.1 给出了有 i 个质点的区间数 n_i. 我们要用试验数据拟合一个 Poisson 分布 $P(\lambda)$.

表 5.1 α 质点散射试验数据

i	n_i	$n\pi_i^*$	i	n_i		$n\pi_i^*$
0	57	54.399	7	139		140.325
1	203	210.523	8	45		67.822
2	383	407.361	9	27		29.189
3	525	525.496	10	10		
4	532	508.418	11	4	16	17.075
5	408	393.515	12	2		
6	273	253.817	和 n	2608		2608

为此, 先将 X 的值域合并为 $k = 11$ 个组, $0, 1, \cdots, 9$ 各为一组, 10 及 10 以上合并为一组. 令 $n_\sigma = \sum_{i=10}^{\infty} n_i$, $\pi_i(\lambda) = \mathrm{e}^{-\lambda}\lambda^i/i!$, 则似然方程为

$$\sum_{i=0}^{9} n_i(i/\lambda - 1) + \sum_{i=10}^{\infty}(i/\lambda - 1)\pi_i(\lambda) \Big/ \sum_{i=10}^{\infty} \pi_i(\lambda) = 0,$$

即 $\lambda = n^{-1}\big\{\sum_{i=0}^{9} in_i + n_\sigma \sum_{i=10}^{\infty} i\pi_i(\lambda)/\sum_{i=10}^{\infty}\pi_i(\lambda)\big\}$, 因而似然方程的解 $\hat{\lambda}$ 可用迭代法求得. 此处我们采用它的一个近似解 $\lambda^* = \bar{X} = \sum_{i=0}^{12} in_i/n = 3.870$. 再用公式 $\pi_i^* = \mathrm{e}^{-3.870}3.870^i/i!$ $(i = 0, 1, \cdots, 9)$, $\pi_{10}^* = 1 - \sum_{i=0}^{9}\pi_i^*$ 算出 π_i^* 和 $n\pi_i^*$, 进而求得 K_n 的具体数值为 $\big\{\sum_{i=0}^{9} n_i^2/(n\pi_i^*) + n_\sigma^2/(n\pi_{10}^*)\big\} - n = 12.885$. 因在数据来自Poisson分布的原假设下, $K_n \rightsquigarrow \chi_9^2$, 故拟合优度为 $P\{\chi_9^2 > 12.885\} \approx 0.17$. 由这一准则, 人们可预期在理想的条件下, 在 100 个类似的情形中, 大约有 17 个的拟合程度比这里的数据要差.

在这个例子中, 我们将后面频数偏小的三组数据合并为一组, 为的是加大这组的经验频数, 以减少因有关组别频数过小对分布逼近造成的不利影响. 但由此会造成两个问题. 其一是按照定理 5.32 和定理 5.33, 分组应与样本无关, 否则极限分布就会有些许偏差. 但实际上人们大多根据样本的值去分组, 以使各组频数都不是太小且组间差距不悬殊. 好在从实用上看由此造成的影响也并没有那么重要. 其二是并组后加大了估计未知参数的复杂程度. 例如在本例中, 若不并组, 则 λ 的 MLE 即为 $\bar{X}$, 并组后要用迭代法才能得到似然方程的数值解. 在实用中一般仍用未分组时的 MLE

估计参数. 这也会对极限分布有些影响, 但其实际意义并不大.

4. 列联表用于独立性检验

作为 χ^2 检验的一个应用, 现在介绍如何用列联表检验两个属性变量的独立性.

设在所考察的总体中, 每个个体都有 A, B 两种属性, 属性 A 有 r 个级别 $A_1, \cdots, A_r$, 属性 B 有 s 个级别 $B_1, \cdots, B_s$. 比如要考察农民的受教育水平与性别是否有关联, 可以把学历分成若干级别, 而人的性别则有两个级别.

今从该总体中独立随机地抽取 n 个个体, 以 n_{ij} 记属性为 (A_i, B_j) 的个体数, 表 5.2 列出了有关数据. 这样的表格称为 $r \times s$ 列联表.

为要检验 "属性 A, B 独立" 的假设, 我们可以引进一个二维随机向量 (X, Y). 若随机抽取的个体属于 (A_i, B_j), 则定义 $X = i, Y = j$, $i = 1, \cdots, r, j = 1, \cdots, s$. 所要检验的假设为

$$H : X, Y \text{ 独立}, \tag{5.62}$$

即存在非负常数 $p_{i\cdot}, i = 1, \cdots, r$ 及 $p_{\cdot j}, j = 1, \cdots, s$, $\sum_{i=1}^{r} p_{i\cdot} = \sum_{j=1}^{s} p_{\cdot j} = 1$, 使 $P(X = i, Y = j) = p_{i\cdot} p_{\cdot j}, i = 1, \cdots, r, j = 1, \cdots, s$. 这就把这个问题归结为 (5.57) 的检验问题. 由于 $r + s$ 个参数有两个约束, 故自由参数的个数为 $r + s - 2$.

由于各种应用的需要, 列联表分析已成为统计方法研究的一个重要方面. 一般可以考虑个体属于 (A_i, B_j) 的概率 p_{ij} 的变化规律, 而与 log 线性模型有着内在的密切联系. 在统计学中列联表分析属于离散多元分析的范畴.

为用 χ^2 方法检验假设 (5.62), 先要给出 $p_{i\cdot}$ 和 $p_{\cdot j}$ 的 MLE, 即在约束 $\sum_{i=1}^{r} p_{i\cdot} = \sum_{j=1}^{s} p_{\cdot j} = 1$ 之下, 使 $\sum_{i=1}^{r} \sum_{j=1}^{s} n_{ij} \log(p_{i\cdot} p_{\cdot j}) \equiv \sum_{i=1}^{r} n_{i\cdot} \log p_{i\cdot} + \sum_{j=1}^{s} n_{\cdot j} \log p_{\cdot j}$ 达到最大, 不难得到相应的估计为

$$\hat{p}_{i\cdot} = n_{i\cdot}/n, \quad \hat{p}_{\cdot j} = n_{\cdot j}/n, \quad i = 1, \cdots, r, \quad j = 1, \cdots, s,$$

其中 $n_{i\cdot} = \sum_{j=1}^{s} n_{ij}, n_{\cdot j} = \sum_{i=1}^{r} n_{ij}$, 而 (A_i, B_j) 组的理论频数估计值为 $n\hat{p}_{i\cdot}\hat{p}_{\cdot j} = n_{i\cdot} n_{\cdot j}/n$, 由此得到 χ^2 统计量

$$K_n = \sum_{i=1}^{r} \sum_{j=1}^{s} \frac{(n n_{ij} - n_{i\cdot} n_{\cdot j})^2}{n n_{i\cdot} n_{\cdot j}} \equiv \sum_{i=1}^{r} \sum_{j=1}^{s} \frac{(n_{ij} - n\hat{p}_{i\cdot}\hat{p}_{\cdot j})^2}{n\hat{p}_{i\cdot}\hat{p}_{\cdot j}}. \tag{5.63}$$

由定理 5.33, 若假设 H 成立, 则当 $n\to\infty$, $K_n \rightsquigarrow \chi^2_{(r-1)(s-1)}$. 基于这个极限分布, 即可计算出数据的拟合优度, 或者按指定的水平作出大样本检验.

5. 列联表用于齐一性检验

如果检验问题是要检验若干个分布 "齐一" 即一致的假设, 则称为**齐一性检验**. 这类问题一般可表述为: 有 r 个总体, 都只取 $a_1,\cdots,a_s$ 这 s 个不同的值, 第 i 个总体的变量记为 X_i. 从各个总体独立随机抽样共 n 次, 第 i 个总体抽取了 n_i 次, 其中取值 a_j 的为 n_{ij} 个. 与前面独立性检验问题不同的是, 此处的 n_i (相当于那儿的 $n_{i\cdot}$) 是给定的而不是随机的. 对齐一性检验我们首先考虑公共分布完全未知的情形, 即所要检验的假设为

$$H: P(X_i=a_j)=p_j>0,\ j=1,\cdots,s,\ i=1,\cdots,r. \tag{5.64}$$

沿用前面的列联表, 其中的 $n_{i\cdot}$ 用 n_i 代替. 将每个样本的 χ^2 值相加得总的 χ^2 值

$$\chi^2=\sum_{i=1}^{r}\sum_{j=1}^{s}\frac{(n_{ij}-n_ip_j)^2}{n_ip_j}. \tag{5.65}$$

但此处的 $p_1,\cdots,p_s$ 未知, 应由合样本来估计. 合样本的 log 似然函数为

$$\sum_{j=1}^{s-1} n_{\cdot j}\log p_j + n_{\cdot s}\log\big(1-\sum_{1}^{s-1}p_j\big),$$

由此即得 p_j 的 MLE 为 $p_j^*=n_{\cdot j}/n=\sum_{l=1}^{r}n_{lj}/n, j=1,\cdots,s.$ 以此代入 (5.65) 即得 χ^2 检验统计量

$$K_n=\sum_{i=1}^{r}\sum_{j=1}^{s}\frac{(n_{ij}-n_ip_j^*)^2}{n_ip_j^*}\equiv\sum_{i=1}^{r}\sum_{j=1}^{s}\frac{(nn_{ij}-n_in_{\cdot j})^2}{nn_in_{\cdot j}}. \tag{5.66}$$

定理 5.34 *当 r 个总体的真参数均为 $p_1>0,\cdots,p_s>0$ 时, 有*

$$K_n\rightsquigarrow\chi^2_{(r-1)(s-1)},\ n_{\min}:=\min\{n_1,\cdots,n_r\}\to\infty.$$

由此可见, 在这种情况下得到的 χ^2 统计量不仅与独立性检验中的类似, 而且两者的极限分布也完全一样. 但需要指出的是, 这一定理不是定理 5.33 的一个推论, 我们必须另行证明.

在证明定理之前, 我们先简要介绍一下矩阵的 Kroneck 乘积的有关知识. 设矩阵 $A=(a_{ij})\in\mathbf{C}^{m\times n}, B=(b_{st})\in\mathbf{C}^{p\times q}$ 为任意的两个复数阵. 称如下的分块矩阵

$$A\otimes B=\begin{pmatrix} a_{11}B & a_{12}B & \cdots & a_{1n}B\\ a_{21}B & a_{22}B & \cdots & a_{2n}B\\ \vdots & \vdots & & \vdots\\ a_{m1}B & a_{m2}B & \cdots & a_{mn}B\end{pmatrix}\in\mathbf{C}^{mp\times nq}$$

是 A 与 B 的 **Kronecker 乘积**, 或**直积**, 或**张量积**. 关于 Kronecker 乘积的有关性质, 可参看关于矩阵理论或多元分析的有关专著或参考书, 中文的例如倪国熙 (1984). 此

处我们只用到乘法的结合律以及以下三条:

1° $(A\otimes B)^\tau=(A^\tau)\otimes(B^\tau)$.

2° $(A_1\otimes B_1)(A_2\otimes B_2)=(A_1A_2)\otimes(B_1B_2)$.

3° $\mathrm{tr}(A\otimes B)=(\mathrm{tr}A)(\mathrm{tr}B), A\in\mathbf{C}^{m\times m}, B\in\mathbf{C}^{n\times n}$.

证明 (1) 记 $\xi_{nij}=(n_{ij}-n_ip_j)/\sqrt{n_ip_j}, j=1,\cdots,s, i=1,\cdots,r$, 则

$$\frac{n_{nij}-n_ip_j^*}{\sqrt{n_ip_j^*}}=\xi_{nij}-\frac{\sqrt{n_i}(p_j^*-p_j)}{\sqrt{p}_j}+\frac{n_{ij}-n_ip_j^*}{\sqrt{n}_i}\Big(\frac{1}{\sqrt{p_j^*}}-\frac{1}{\sqrt{p}_j}\Big). \tag{5.67}$$

因 $\mathrm{E}(p_j^*)=p_j>0, \mathrm{Var}\,(p_j^*)=p_j(1-p_j)/n$, 故 $1/\sqrt{p_j^*}-1/\sqrt{p_j}=o_p(1)$,

$$\frac{n_{ij}-n_ip_j^*}{\sqrt{n}_i}=\frac{n_{ij}-n_ip_j}{\sqrt{n}_i}-\sqrt{n_i}(p_j^*-p_j)=O_p(1)+O_p(1)=O_p(1),$$

因而 (5.67) 右边第三项为 $o_p(1)$. 将 $p_j^*=\sum_{l=1}^r n_{lj}/n$ 代入 (5.67) 并化简, 即得

$$\frac{n_{ij}-n_ip_j^*}{\sqrt{n_ip_j^*}}=\xi_{nij}-\sqrt{n_i/n}\sum_{l=1}^r\sqrt{n_l/n}\xi_{nlj}+o_p(1):=\eta_{nij}+o_p(1), \tag{5.68}$$

故由 (5.66) 及 (5.68),

$$K_n=\sum_{i=1}^r\sum_{j=1}^s\eta_{nij}^2+o_p(1),\ n_{\min}\to\infty. \tag{5.69}$$

(2) 记 $\xi_{ni}=(\xi_{ni1},\cdots,\xi_{nis})^\tau: s\times1$, $\xi_n=(\xi_{n1}^\tau,\cdots,\xi_{nr}^\tau)^\tau: rs\times1$. 对于随机向量 $\eta_{ni}: s\times1$ 和 $\eta_n: rs\times1$, 我们也可如法定义. 记 $\psi=(\sqrt{p_1},\cdots,\sqrt{p_s})^\tau: s\times1$, $\phi_n=(\sqrt{n_1/n},\cdots,\sqrt{n_r/n})^\tau: r\times1$.

因 ψ 为 s-维单位向量, 故存在 $s\times(s-1)$ 矩阵 A, 使 $(\psi\ A)$ 为 s 阶正交阵. 记 $\zeta_{ni}=A^\tau\xi_{ni}: (s-1)\times1$, $\zeta_n=(\zeta_{n1}^\tau,\cdots,\zeta_{nr}^\tau)^\tau: r(s-1)\times1$. 一如定理 5.32 证明中所示, 由 $\psi^\tau\xi_{ni}=0$ 可得 $\xi_{ni}=(\psi\psi^\tau+AA^\tau)\xi_{ni}=A\zeta_{ni}$, $\xi_n=(I_r\otimes A)\zeta_n$, $\eta_n=[(I_r-\phi_n\phi_n^\tau)\otimes I_s]\xi_n$, 且由 $\xi_{ni}\rightsquigarrow N(0,I_s-\psi\psi^\tau)$ 及 $\psi^\tau A=0$, $A^\tau A=I_{s-1}$, 有

$$\zeta_{ni}\rightsquigarrow N(0,I_{s-1}),\ \zeta_n\rightsquigarrow N(0,I_{r(s-1)}). \tag{5.70}$$

我们有

$$\begin{aligned}\sum_{i=1}^r\sum_{j=1}^s\eta_{nij}^2&=\eta_n^\tau\eta_n=\xi_n^\tau[(I_r-\phi_n\phi_n^\tau)\otimes I_s]\xi_n\\&=\zeta_n^\tau[I_r\otimes A^\tau][(I_r-\phi_n\phi_n^\tau)\otimes I_s][I_r\otimes A]\zeta_n\\&=\zeta_n^\tau[(I_r-\phi_n\phi_n^\tau)\otimes A^\tau A]\zeta_n\\&=\zeta_n^\tau[(I_r-\phi_n\phi_n^\tau)\otimes I_{s-1}]\zeta_n:=\zeta_n^\tau M_n\zeta_n,\end{aligned} \tag{5.71}$$

因 M_n 为 $r(s-1)$ 阶对称幂等阵, 且 $\mathrm{tr}(M_n)=(r-1)(s-1)$, 故由 (5.69)—(5.71), 即得 $K_n\rightsquigarrow\chi^2_{(r-1)(s-1)}$. ■

在公共分布部分未知的一般情形, 所要检验的假设为

$$H: P(X_i=a_j)=p_j(\theta),\ j=1,\cdots,s,\ i=1,\cdots,r,$$

此处 $\theta\in\Theta$, Θ 为 $\mathbf{R}^t$ 的一个开子集, $t\leqslant s-1$. 当 $t<s-1$ 时, 未知参数的 MLE 没有上面的显式表达, 但仍可证明, 基于这些估计如法构作的 χ^2 统计量 K_n,

当 $\min\{n_1,\cdots,n_r\}\to\infty$ 时, 在 H 成立时的极限分布仍为 χ^2 分布, 其自由度为 $r(s-1)-t$. 当 $t=s-1$ 时极限分布与定理 5.34 中的一致.

6. 其他检验

除了 χ^2 拟合优度检验外, 还有一些检验方法, 可用于检验形如 (5.56) 或 (5.57) 的假设. 一些方法跟 χ^2 方法一样, 对理论分布的形式没有特殊的限定. 这些方法大体可分为两类, 其一是 χ^2 方法的变形或进一步的发展, 另一类则是**经验过程**的某个泛函. 著名的 Kolmogorov 检验即为后者的典型代表. 我们在 1.4.1 节例 1.5 中曾有过简单的介绍. 有些方法则是针对某个特定分布族的, 例如检验一组数据是否来自某个正态总体的 Wilk-Shapiro 的 W 检验和 D'Agostino 的 D 检验. 关于这些检验以及拟合优度检验的系统论述, 读者可参看杨振海, 程维虎, 张军舰 (2011), 以及陈家鼎, 孙山泽, 李东风, 刘力平 (2006).

(1) Kolmogorov 检验 设 $X_1,\cdots,X_n$ 为取自某个 k-维总体 X 的 i.i.d. 样本, 问题是要检验假设 (5.56):

$$H: X_1,\cdots,X_n \text{ 的公共分布为 } F,$$

其中 F 为一完全已知的分布.

以 ξ 记一随机变量, 它取值每个 X_i 的概率均为 $1/n$, 则 ξ 的分布函数 $\mathbb{F}_n(x)\equiv\mathbb{F}_n(x;X_1,\cdots,X_n)$ 称为样本 $X_1,\cdots,X_n$ 的**经验分布函数**. 若原假设 H 成立, 则由 Glivenko-Cantelli 定理,

$$D_n:=\sup_x|\mathbb{F}_n(x)-F(x)|\to 0 \ \text{ a.s., 当 } \ n\to\infty.$$

因此, 为检验假设 (5.56), 选取某个 c, 以 $\{D_n>c\}$ 为否定域的检验是一个合理的检验. 问题在于若 D_n 的分布与 F 有关, 要对每个特定的 F 确定临界值 c 不是一件容易的事. 好在**当 F 为连续的一维分布函数时, D_n 的分布与 F 无关**. 因此, 对一特定的 F 及水平 α 确定的临界值 c, 就是对这一类的 F 全都适用的. 这个检验所以称为 Kolmogorov 检验, 是由于检验统计量的极限分布是 Kolmogorov (1933) 建立的.

为了证明 D_n 的分布与 F 无关, 以 $X_{(1)}\leqslant\cdots\leqslant X_{(n)}$ 记 $X_1,\cdots,X_n$ 的次序统计量, 则经验分布函数 $\mathbb{F}_n(x)$ 有以下表达式

$$\mathbb{F}_n(x)=\sum_{i=1}^n I(X_i\leqslant x)=\begin{cases}0, & x<X_{(1)},\\ k/n, & X_{(k)}\leqslant x<X_{(k+1)}, k=1,\cdots,n-1,\\ 1, & x\geqslant X_{(n)}.\end{cases}\tag{5.72}$$

我们有

$$\begin{aligned}&\sup\{|\mathbb{F}_n(x)-F(x)|: x<X_{(1)}\}=F(X_{(1)}),\\ &\sup\{|\mathbb{F}_n(x)-F(x)|: X_{(k)}\leqslant x<X_{(k+1)}\}\\ =&\max\{|k/n-F(X_{(k)})|,|F(X_{(k+1)})-k/n|\},\quad 1\leqslant k\leqslant n-1,\\ &\sup\{|\mathbb{F}_n(x)-F(x)|: x\geqslant X_{(n)}\}=1-F(X_{(n)}).\end{aligned}$$

因 F 连续, 故 $\{U_{(k)}:=F(X_{(k)}), k=1,\cdots,n\}$ 可视为取自均匀分布 $U(0,1)$ 的

n 个次序样本, 故

$$\sup_x |\mathbb{F}_n(x) - F(x)| = \max_{1\leqslant k\leqslant n}\{|k/n - U_{(k)}|, |U_{(k)} - (k-1)/n|\}$$

的分布与 F 无关. 关于 Kolmogorov 检验统计量的极限分布, 有如下的定理.

定理 5.35 (Kolmogorov 定理) *若随机变量 X 有连续的一维分布函数 F, 则*

$$\lim_{n\to\infty} P(\sqrt{n}D_n \leqslant x) = Q(x) \equiv 1 + 2\sum_{k=1}^{\infty}(-1)^k \exp\{-2k^2x^2\}, \quad x > 0. \tag{5.73}$$

这个著名的结果有多个证明方法, 其中最简洁而漂亮的证法是基于**经验过程** $\{\mathbb{G}_n(x) := \sqrt{n}\big(\mathbb{F}_n(x) - F(x)\big), x \in [-\infty,\infty]\}$ 的弱收敛理论. 下面我们扼要介绍证明的梗概. 根据上面的讨论, 不妨假定 $X_1,\cdots,X_n$ 为取自均匀分布 $U(0,1)$ 的 i.i.d. 样本, $\mathbb{F}_n$ 为基于此样本的经验分布函数. 此时 $\mathbb{G}_n(t) = \sqrt{n}\big(\mathbb{F}_n(t) - t\big), t \in [0,1]$. 由附录 A.4 中的例 A.3 及定理 A.24, 在 $D[0,1]$ 上有

$$\mathbb{G}_n \rightsquigarrow \mathbb{B}, \quad n \to \infty,$$

其中 $\{\mathbb{B}(t), 0 \leqslant t \leqslant 1\}$ 为一标准的 Brown 桥. 由连续映照定理,

$$\sup_{0\leqslant t\leqslant 1} |\mathbb{G}_n(t)| \rightsquigarrow \sup_{0\leqslant t\leqslant 1} |\mathbb{B}(t)|.$$

剩下的事就是求出上述极限变量的分布函数. 详见 Billingsley (1986) 或 (1999).

利用这个极限分布即可得到 (5.56) 的一个大样本检验: 取定 α, 求得 c_α, 使 $1 - Q(c_\alpha) = \alpha$, 当 $\sqrt{n}D_n > c_\alpha$ 时否定 H. 当 $\alpha = 0.05$ 及 0.01 时, c_α 分别为 1.358 及 1.628.

与 χ^2 检验比较, 在总体具有连续且完全已知的分布函数的情况下, Kolmogorov 检验的优势在于, 它有较大的功效, 没有分组的随意性, 以及统计量的确切分布与极限分布差距较小等. χ^2 检验的优势在于, 这种方法对任何维数都适用, 对理论分布没有什么限制. 特别是在 X 为多维时, 极限分布也不依赖于理论分布 F. 但当 X 为多维时, 定理 5.35 已不再成立, 此时极限分布的形式依赖于 F. 而在理论分布 F 带参数时, 用 χ^2 检验没有什么原则困难, 但 Kolmogorov 检验则不行. 关于这方面更仔细的讨论, 可参看杨振海, 程维虎, 张军舰 (2011).

(2) Cramér-von Mises 检验 这一检验同样是用于检验上述假设 (5.56). 检验统计量为

$$C_n = n\int_{-\infty}^{\infty} \big(\mathbb{F}_n(x) - F(x)\big)^2 dF(x).$$

当 C_n 大于某个 c 时否定 H. 经计算知道, 当 F 为连续的一维分布函数时,

$$C_n = \frac{1}{12n} + \sum_{i=1}^{n}\Big(F(X_{(i)}) - \frac{2i-1}{2n}\Big)^2 = \frac{1}{12n} + \sum_{i=1}^{n}\Big(U_{(i)} - \frac{2i-1}{2n}\Big)^2.$$

此处, $U_i := F(X_i) \sim U(0,1)$, $U_{(1)},\cdots,U_{(n)}$ 为取自 $U(0,1)$ 的 n 个次序样本. 由此可见, C_n 的分布与 F 无关. 1936 年 Smirnov 求得了 $n\to\infty$ 时 C_n 的极限分布的特征函数, Andersen 和 Darling (1952) 通过反演得到了极限分布函数.

定理 5.36 若随机变量 X 有连续的一维分布函数 F, 则

$$C_n \rightsquigarrow \xi = \sum_{k=1}^{\infty} \frac{\zeta_k}{k^2\pi^2} = \frac{1}{6} + \sum_{k=1}^{\infty} \frac{\zeta_k - 1}{k^2\pi^2},$$

其中 $\zeta_1, \zeta_2, \cdots$ 为 i.i.d. χ_1^2 变量, 即

$$\lim_{n\to\infty} P(C_n \leqslant x) = P(\xi \leqslant x), \quad x > 0.$$

有关证明和上式右方的计算, 可参看 Durbin (1973a), van der Vaart (1998) 第 171, 277 和 295 页, 以及杨振海, 程维虎, 张军舰 (2011).

计算表明, 当原假设成立时, C_n 的确切分布收敛于极限分布的速度很快, 例如极限分布的上 0.1, 0.05, 和 0.025 分位点分别为 0.347, 0.461 和 0.743, 而 C_3 的这些分位点精确到三位小数也与此相同. 该检验也与 Kolmogorov 检验有同样的缺点.

(3) 两样本的 Smirnov 检验 经验分布函数也可用于多样本检验问题, 以著名的两样本 Smirnov 检验为例, 设 $X_{i1}, \cdots, X_{in_i}$ 为抽自一维总体 X_i 的 i.i.d. 样本, X_i 有连续的分布函数 F_i, $i = 1, 2$, 且两样本整体独立, 以 $\mathbb{F}_{1n_1}$ 和 $\mathbb{F}_{2n_2}$ 分别记这两组样本的经验分布函数, 令

$$D^+(n_1, n_2) = \sup_x \big(\mathbb{F}_{1n_1}(x) - \mathbb{F}_{2n_2}(x)\big),$$
$$D(n_1, n_2) = \sup_x \big|\mathbb{F}_{1n_1}(x) - \mathbb{F}_{2n_2}(x)\big|,$$

Smirnov 证明了如下的结果.

定理 5.37 (Smirnov) 设 $F_1 = F_2$ 且处处连续, 则当 $n_1 \to \infty, n_2 \to \infty$ 时,

$$P\Big(\sqrt{\frac{n_1 n_2}{n_1 + n_2}} D^+(n_1, n_2) \leqslant x\Big) \to 1 - \exp(-2x^2),\ x > 0,$$
$$P\Big(\sqrt{\frac{n_1 n_2}{n_1 + n_2}} D(n_1, n_2) \leqslant x\Big) \to Q(x),\ x > 0,$$

其中 $Q(x)$ 的定义见 (5.73).

D^+ 和 D 分别称为单向和双向的 **Smirnov 检验统计量**, 分别用于检验单边和双边的假设.

Smirnov 还对一样本情形的 $D_n^+ := \sup_x \big(\mathbb{F}_n(x) - F(x)\big)$ 建立了 $\sqrt{n}D_n^+$ 的极限分布, 其中 F 连续. 张里千 (1955) (1956) 对 Kolmogorov-Smirnov 统计量 $\sqrt{n}D_n$ 和 $\sqrt{n}D_n^+$ 的分布所作的渐近展开, 是国际国内同行公认的重要工作.

5.6 次序统计量

我们已在好几处提到过次序统计量, 这也从一个侧面显示了次序统计量的重要性. 次序统计量渗透到数理统计的多个分支, 它既可用于典型的非参数统计问题, 也可用于典型的参数统计问题, 在诸如水文、地质、风险分析、灾害评估、保险和金融

等多个领域都有重要的应用. 其理论研究跨越数理统计和概率论的多个分支, 取得了令人瞩目的成就.

本节的目的就是介绍次序统计量的初步理论和若干应用, 集中讨论它的分布和收敛性质, 特别注重于有关的渐近分布理论. 这里我们特别指出, **本节涉及的随机变量 X 和分布 F 都是一维的**.

5.6.1 基本概念

设有一组样本 $X_1,\cdots,X_n$, 将其按由小到大的次序排列为

$$X_{(1)}\leqslant X_{(2)}\leqslant\cdots\leqslant X_{(n)}, \tag{5.74}$$

则 $(X_{(1)},\cdots,X_{(n)})$ 称为样本 $X_1,\cdots,X_n$ 的**次序统计量** (order statistics). 习惯上也常把序列 (5.74) 的一部分称为次序统计量. 特别地, $X_{(i)}$ 常称为第 i 个次序统计量. 如果 $X_1,\cdots,X_n$ 是从分布为 F 的总体中抽出的 i.i.d. 样本, 则称 (5.74) 为从 F 中抽出的 (容量为 n) 的**次序样本** (ordered sample). 如果样本中出现 $X_i=X_j, i\neq j$ 的情形, 则称排序出现了 "结" (tie). 可以有多种方式消除结, 例如当出现 $X_i=X_j, i<j$ 的情形, 则按抽样先后将 X_i 排在 X_j 之前. 对 i.i.d 样本而言, 如果总体的分布函数连续, 则有 $P\{X_{(1)}<X_{(2)}<\cdots<X_{(n)}\}=1$.

在涉及不同容量的次序样本时, 序列 (5.74) 常常改写为

$$X_{1:n}\leqslant X_{2:n}\leqslant\cdots\leqslant X_{n:n}.$$

在序列 (5.74) 中, 若 X_i 处于第 R_i 个位置, 则称 X_i 在 $\{X_1,\cdots,X_n\}$ 中的秩 (简称 X_i 的秩) 为 R_i, 即 $R_i=r_i$ 当且仅当 $X_i=X_{(r_i)}$. 此时, $R:=(R_1,\cdots,R_n)$ 及由 R 派生出来的任何统计量都称为**秩统计量** (rank statistics). 当总体分布 F 连续时, 秩统计量 R 以概率 1 唯一确定.

关于秩统计量的分布, 有下面的定理. 秩方法在非参数领域的重要地位, 在很大程度上来源于这一简单事实. 就秩统计量而言, 本书只限于介绍这一定理.

定理 5.38 设 $X_1,\cdots,X_n$ 是从分布函数 F 连续的一维总体中抽取的 i.i.d. 样本, R 为 $(X_1,\cdots,X_n)$ 的秩, 则对 $(1,2,\cdots,n)$ 的任一置换 $r:=(r_1,r_2,\cdots,r_n)$ 有

$$P\big(R=(r_1,r_2,\cdots,r_n)\big)=1/n!.$$

证明 证明可由对称性得出. 事实上, 对 $(1,2,\cdots,n)$ 的任一置换 r, 有

$$\begin{aligned}P\{R=r\}&=P\big\{(X_1,\cdots,X_n)=(X_{(r_1)},\cdots,X_{(r_n)})\big\}\\&=P\big\{(X_{d_1},\cdots,X_{d_n})=(X_{(1)},\cdots,X_{(n)})\big\}=P\{X_{d_1}<\cdots<X_{d_n}\},\end{aligned}$$

其中 $(d_1,\cdots,d_n)$ 是 $(1,2,\cdots,n)$ 的一个置换. 由 $X_1,\cdots,X_n$ i.i.d.,

$$(X_1,\cdots,X_n)\overset{d.}{=}(X_{d_1},\cdots,X_{d_n}),$$

故对任一 r,

$$P\{R=r\}=P\{X_1<\cdots<X_n\},$$

由于这样的置换 r 共有 $n!$ 个, 因而 $P\{R=r\}=1/n!$. 证毕. ■

此前我们已经多次提到样本 $X_1,\cdots,X_n$ 的**经验分布函数**

$$\mathbb{F}_n(x)=\frac{1}{n}\sum_{i=1}^{n}I(X_i\leqslant x),\ x\in\mathbf{R}^1,$$

参看 (5.72). 今设 F 为一右连续的分布函数. 定义 F 的**分位函数**为

$$F^{-1}(t):=\inf\{x:F(x)\geqslant t\},\ 0<t<1. \tag{5.75}$$

易见, 这样定义的分位函数是一个左连续函数. 对经验分布函数 $\mathbb{F}_n$, 自然可同样定义其分位函数 $\mathbb{F}_n^{-1}:=\inf\{x:\mathbb{F}_n(x)\geqslant t\}$, 称之为**经验分位函数**. 当分布函数 F 连续时, 无需考虑 "结" 的问题, 于是有

$$\mathbb{F}_n^{-1}(t)=X_{i:n},\quad \text{若 } (i-1)/n<t\leqslant i/n,$$

即

$$\mathbb{F}_n^{-1}(t)=X_{\langle nt\rangle:n},\ \ 0<t<1, \tag{5.76}$$

其中 $\langle nt\rangle$ 表示不小于 nt 的最小整数.

与总体分布有关的一个重要概念是**分位数**的概念. 设一维随机变量 X 有分布函数 F, $0<p<1$, 则 X (或分布 F) 的 p-**分位数**定义为满足

$$P(X<\nu)\leqslant p\leqslant P(X\leqslant\nu),\ \text{即 } F(\nu-0)\leqslant p\leqslant F(\nu)$$

的任一实数 ν, 记为 ξ_p. 其中 $\xi_{1/2}$ 特别重要, 称为 X 或 F 的中位数, 记为 $\mathrm{med}(X)$. 关于分位数的一些基本性质, 可参看 1.4.1 节例 1.4, 此处不再重复. 根据这个定义, $F^{-1}(p)$ 即为 F 最小的 p-分位数. 类似地, **样本 p-分位数** $\hat{\xi}_{n,p}$ 可以定义为满足

$$\mathbb{F}_n(\hat{\xi}_{n,p}-)\leqslant p\leqslant\mathbb{F}_n(\hat{\xi}_{n,p})$$

的任一随机变量. 同样, $\mathbb{F}_n^{-1}(p)=X_{\langle np\rangle:n}$ 即为最小的样本 p-分位数. 在有些文献中, 样本 p-分位数定义为 $X_{[np]:n}$ 和 $X_{[np]+1:n}$ 的如下的凸组合:

$$\hat{\xi}_{n,p}:=X_{[np]:n}+\Big\{(n+1)\Big(p-\frac{[np]}{n+1}\Big)\Big\}^*\big(X_{[np]+1:n}-X_{[np]:n}\big),$$

此处, $[x]$ 表示不超过 x 的最大整数, $a^*:=\min(a,1)$. 此时样本中位数定义为

$$\hat{\xi}_{n,1/2}=\begin{cases}(X_{n/2:n}+X_{n/2+1:n})/2, & n\text{ 为偶数},\\ X_{(n+1)/2:n}, & n\text{ 为奇数},\end{cases}$$

当 n 较大时, 样本 p-分位数也可取为 $X_{[(n+1)p]:n}$ 或 $X_{[np]:n}$. 易见, 它们都与次序统计量有关.

如果总体 p-分位数唯一, 则可用次序统计量或样本 p-分位数来估计, **这是次序统计量的一项重要应用**. 另外, 在例 1.4 中我们还将 ξ_p 的 M-估计定义为极值问题

$$\sum_{i=1}^{n}\rho_p(X_i-t)=\min$$

的任一可测解, 此处 $\rho_p(z)=z[p-I(z<0)]$ 为一凸函数.

以后将会看到, 当考虑大样本问题时, 为估计总体的 p-分位数, 无论采用上述哪种估计量, 都没有多大差别, 它们都有相同的极限分布.

1. 次序统计量的分布

今设 $X_1,\cdots,X_n$ 是从分布 F 中抽取的 i.i.d. 样本, 并由此出发讨论有关的次序统计量的分布问题. 以 F_r 记 $X_{(r)}$ 的分布函数, 则

$$\begin{aligned}F_r(x)&=P(X_{(r)}\leqslant x)=P\big(X_1,\cdots,X_n\text{ 中至少有 }r\text{ 个}\leqslant x\big)\\&=\sum_{j=r}^{n}P\big(X_1,\cdots,X_n\text{ 中恰有 }j\text{ 个}\leqslant x\big)\\&=\sum_{j=r}^{n}\binom{n}{j}F^j(x)(1-F)^{n-j}(x)\\&=\frac{n!}{(r-1)!(n-r)!}\int_0^{F(x)}t^{r-1}(1-t)^{n-r}dt.\end{aligned}$$

最后一个等式是基于恒等式

$$\sum_{j=r}^{n}\binom{n}{j}p^j(1-p)^{n-j}\equiv\frac{n!}{(r-1)!(n-r)!}\int_0^p t^{r-1}(1-t)^{n-r}dt,\ 0\leqslant p\leqslant 1.$$

若 F 有密度 f, 则 F_r 也有密度 f_r, 且

$$f_r(x)=\frac{n!}{(r-1)!(n-r)!}F(x)^{r-1}(1-F)^{n-r}(x)f(x).$$

在次序统计量中,

$$X_{(n)}=\max\{X_1,\cdots,X_n\}$$

和

$$X_{(1)}=\min\{X_1,\cdots,X_n\}$$

有着特殊的地位, 分别称为**样本极大值**和**样本极小值**, 统称为**样本极值**. 我们也不难写出它们的分布. 例如 $X_{(n)}$ 有分布

$$F_n(x)=F^n(x),\ f_n(x)=nF^{n-1}(x)f(x).$$

设 F 有密度 f 且 $1\leqslant r<s\leqslant n$. 不难证明 $(X_{(r)},X_{(s)})$ 的联合密度 $f_{rs}(x,y)$ 为

$$\frac{n!}{(r-1)!(s-r-1)!(n-s)!}F^{r-1}(x)\big(F(y)-F(x)\big)^{s-r-1}(1-F)^{n-s}(y)f(x)f(y),$$

或 0, 视 $x<y$ 或 $x\geqslant y$ 而定. 而 $(X_{(1)},\cdots,X_{(n)})$ 的密度函数为

$$f_{12\cdots n}(x_1,\cdots,x_n)=\begin{cases}n!f(x_1)f(x_2)\cdots f(x_n),&\text{若 }x_1<x_2<\cdots<x_n,\\0,&\text{否则.}\end{cases}$$

2. 总体分布为 $U(0,1)$ 的情形

当 F 为均匀分布 $U(0,1)$ 即总体密度 $f(x)=I_{(0,1)}(x)$ 时, 上一段的公式取很简单的形式. 此时 $X_{(r)}$ 的密度函数为

$$f_r(x)=\frac{n!}{(r-1)!(n-r)!}x^{r-1}(1-x)^{n-r}I_{(0,1)}(x).$$

$(X_{(r)}, X_{(s)})$ 的联合密度 $f_{rs}(x,y)$ 为

$$\frac{n!}{(r-1)!(s-r-1)!(n-s)!}x^{r-1}(y-x)^{s-r-1}(1-y)^{n-s}I(0<x<y<1).$$

$(X_{(1)},\cdots,X_{(n)})$ 的密度函数则为

$$f_{12\cdots n}(x_1,\cdots,x_n)=\begin{cases}n!, & \text{若 } 0<x_1<x_2<\cdots<x_n<1,\\ 0, & \text{否则.}\end{cases}$$

均匀分布 $U(0,1)$ 的重要性基于以下的两个定理. 下一定理为我们通过均匀分布构造具有给定分布 F 的随机变量, 以及通过随机数构造具有分布 F 的独立样本提供了理论依据.

定理 5.39 (分位数变换) 设 F 为一分布函数, $U\sim U(0,1)$, 则 $F^{-1}(U)\sim F$, 其中 $F^{-1}(u):=\inf\{x: F(x)\geqslant u\}$ 为定义于 $(0,1)$ 区间内的分位函数. 事实上, $F^{-1}(U)\leqslant x\Longleftrightarrow U\leqslant F(x), x\in\mathbf{R}$.

证明 对任一 $u\in(0,1)$, 记 $A(u)=\{x: F(x)\geqslant u\}$.

(i) 先设 $x>F^{-1}(u)$, 则 $x\in A(u)$, 故 $F(x)\geqslant u$. 由 F 的右连续性, 有 $F\big(F^{-1}(u)\big)=\lim_{x\downarrow F^{-1}(u)}F(x)\geqslant u$. 故 $x\geqslant F^{-1}(u)\Longrightarrow F(x)\geqslant u$.

(ii) 反之, 若 $F(x)\geqslant u$, 则 $x\in A(u)$, $x\geqslant F^{-1}(u)$.

因此, $F^{-1}(u)\leqslant x\Longleftrightarrow u\leqslant F(x)$. 故当 $0<F(x)<1$ 时,

$$P\big(F^{-1}(U)\leqslant x\big)=P\big(U\leqslant F(x)\big)=F(x).$$

这就证明了定理的结论. ■

由此不难得出下面的推论.

推论 5.40 设 F 为一右连续的分布函数, $0<u<1$, 则

$$F\big(F^{-1}(u)-\big)\leqslant u\leqslant F\big(F^{-1}(u)\big),$$

且对任一 $y\in(0,1)$, 有

$$P\{F(X)\leqslant y\}\leqslant y\leqslant P\{F(X-0)<y\}.$$

事实上, 由上面的 (i), 对任一 $u\in(0,1)$, 有 $F\big(F^{-1}(u)\big)\geqslant u$. 另一方面, 设 $x<F^{-1}(u)$, 则 $F(x)<u$, 故 $F\big(F^{-1}(u)-\big)=\lim_{x\uparrow F^{-1}(u)}F(x)\leqslant u$. 由此即得上面的第一式. 为证第二式, 设 $U\sim U(0,1)$, 则 $X:=F^{-1}(U)\sim F$. 由第一式, $F(X-0)\leqslant U\leqslant F(X)$, 故对任一 $y\in(0,1)$, 有

$$P\{F(X)\leqslant y\}\leqslant P\{U\leqslant y\}=y=P\{U<y\}\leqslant P\{F(X-0)<y\}.$$

由上面的定理和推论, 立得如下定理.

定理 5.41 设随机变量 X 具有连续的分布函数 F, 则 $F(X)\sim U(0,1)$.

由此可知, 若总体 X 的分布函数 F 连续, $(X_{(1)},\cdots,X_{(n)})$ 为 X 的次序样本, 则 $\big(F(X_{(1)}),\cdots,F(X_{(n)})\big)$ 的分布就是从 $U(0,1)$ 的总体抽取的次序样本的分布. 由于这样的结果, 在某些非参数统计理论问题中, 可以只考虑总体分布为 $U(0,1)$ 的情况. 因而均匀分布在非参数统计中具有特殊的地位.

3. 次序统计量的条件分布

设一维总体 X 具有连续的分布函数 F, $X_{(1)} \leqslant \cdots \leqslant X_{(n)}$ 是从该总体中抽取的次序样本. 要在给定其中的一个, 例如 $X_{(r)}$ 的条件下, 求其余 $n-1$ 个的条件分布. 为此, 设给定 $X_{(r)} = t$. 记

$$F^*(x) = \begin{cases} \big(F(x) - F(t)\big)/\big(1 - F(t)\big), & \text{当 } x > t, \\ 0, & \text{当 } x \leqslant t. \end{cases}$$

$$F_*(x) = \begin{cases} F(x)/F(t), & \text{当 } x < t, \\ 1, & \text{当 } x \geqslant t, \end{cases}$$

则 F^* 和 F_* 分别是在给定 $X \geqslant t$ 和 $X \leqslant t$ 的条件下, X 的条件分布. 由 F 的连续性, $(X_{(1)}, \cdots, X_{(n)})$ 和$X_{(r)}$ 的概率元分别是

$$g\big(x_{(1)}, \cdots, x_{(n)}\big) := n! I\big(x_{(1)} < \cdots < x_{(n)}\big) dF(x_{(1)}) \cdots dF(x_{(n)})$$

以及

$$h(x) := \frac{n!}{(r-1)!(n-r)!} F^{r-1}(x)(1-F)^{n-r}(x) dF(x),$$

故当给定 $X_{(r)} = t$ 时, $(X_{(1)}, \cdots, X_{(r-1)}, X_{(r+1)}, \cdots, X_{(n)})$ 的条件概率元为

$$\begin{aligned} & g\big(x_{(1)}, \cdots, x_{(r-1)}, t, x_{(r+1)}, \cdots, x_{(n)}\big)/h(t) \\ = & \big\{(r-1)! dF_*(x_{(1)}) \cdots dF_*(x_{(r-1)}) I\big(x_{(1)} < \cdots < x_{(r-1)} < t\big)\big\} \\ & \times \big\{(n-r)! dF^*(x_{(r+1)}) \cdots dF^*(x_{(n)}) I\big(t < x_{(r+1)} < \cdots < x_{(n)}\big)\big\}. \end{aligned}$$

由此即得以下定理和推论.

定理 5.42 设 $X_{(1)} < \cdots < X_{(n)}$ 是从一维连续分布 F 中抽取的次序样本, 则在给定 $X_{(r)} = t$ 时, $(X_{(1)}, \cdots, X_{(r-1)}, X_{(r+1)}, \cdots, X_{(n)})$ 的条件分布由以下三个性质确定:

1° $(X_{(1)}, \cdots, X_{(r-1)})$ 与 $(X_{(r+1)}, \cdots, X_{(n)})$ 条件独立;

2° $(X_{(1)}, \cdots, X_{(r-1)})$ 的条件分布即为从截断分布 F_* 中抽取的容量为 $r-1$ 的次序样本的分布;

3° $(X_{(r+1)}, \cdots, X_{(n)})$ 的条件分布即为从截断分布 F^* 中抽取的容量为 $n-r$ 的次序样本的分布.

推论 5.43 (Markov 性质) 在定理条件下, $X_{r:n}$ 在给定 $X_{1:n} = x_1, \cdots, X_{r-1:n} = x_{r-1}$ 时的条件分布, 即为$X_{r:n}$ 在给定 $X_{r-1:n} = x_{r-1}$ 时的条件分布. 因此, 序列 $X_{1:n}, \cdots, X_{n:n}$ 具有 Markov 性质.

5.6.2 次序统计量的极限分布

设 $X_1, \cdots, X_n$ 为取自总体 $X \sim F$ 的 i.i.d. 样本, $(X_{1:n}, \cdots, X_{n:n})$ 为其次序统计量. 次序统计量 $X_{r(n):n}$ 何时以及存在何种渐近分布, 与比值 $r(n)/n$ 在 $n \to \infty$ 时的性状有关, 文献中常依此对次序统计量的序列分类. 本节只讨论以下两类:

1° 若存在 $p\in(0,1)$, 使 $r(n)/n=p+o(n^{-1/2})$, 则 $X_{r(n):n}$ 称为**正则中心项序列**.

2° 若 $r(n)$ 或 $n-r(n)$ 不随 n 变化, 则 $X_{r(n):n}$ 称为**固定边项序列**. 本节只限于介绍 $r(n)=1$ 以及 $r(n)=n$ 的情形, 即**样本极值**的情形.

近几十年来, 对种种情况下的序列 $\{X_{r(n):n}\}$ 的渐近分布的研究, 获得了许多深入的结果. 但对统计应用而言, 最重要的当推以上两种情形. 其他情形可参看有关文献, 例如 Reiss (1989).

次序统计量或样本分位数均可用于总体分位数的估计. 首先讨论这个估计的强相合性, 我们有下面的定理.

定理 5.44 设 $0<p<1$, 分布 F 的 p 分位数 ζ_0 唯一, 且 $r(n)/n=p+o(1)$, 则

$$X_{r(n):n}\to\zeta_0 \quad \text{a.s.}, \quad n\to\infty,$$

即 $X_{r(n):n}$ 是总体 p 分位数 ζ_0 的强相合估计.

证明 任给 $\epsilon>0$, 由 p 分位数 ζ_0 唯一, 知 $F(\zeta_0-\epsilon)<p<F(\zeta_0+\epsilon)$. 取 $\delta>0$ 使 $2\delta<\min\{F(\zeta_0+\epsilon)-p, p-F(\zeta_0-\epsilon)\}$. 由强大数定律, 以概率 1, 当 n 充分大时有

$$\sum_{i=1}^{n} I(X_i\leqslant\zeta_0-\epsilon)<n(p-\delta), \quad \sum_{i=1}^{n} I(X_i\leqslant\zeta_0+\epsilon)>n(p+\delta),$$

且 $n(p-\delta)<r(n)<n(p+\delta)$, 因而 $\zeta_0-\epsilon<X_{r(n):n}<\zeta_0+\epsilon$. 由 ϵ 的任意性, 就证明了定理的结论. ■

1. 正则中心项序列的中心极限定理

下面讨论正则中心项序列的中心极限定理. 为此, 我们先引进下面的引理.

引理 5.45 (正则中心项序列的 Bahadur 表示) 设 $0<p<1$, F 在其 p 分位数 ζ_0 处有导数 $f(\zeta_0)>0$, 且 $r(n)/n=p+o(n^{-1/2})$, 则当 $n\to\infty$ 时有

$$\begin{aligned}\sqrt{n}(X_{r(n):n}-\zeta_0)&=-\sqrt{n}\big(\mathbb{F}_n(\zeta_0)-p\big)/f(\zeta_0)+o_p(1)\\&=-\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\frac{I(X_i\leqslant\zeta_0)-p}{f(\zeta_0)}+o_p(1).\end{aligned}\tag{5.77}$$

证明 设 $U_1,\cdots,U_n$ 为取自均匀分布 $R(0,1)$ 的 i.i.d. 样本, 则其次序统计量 $U_{1:n},\cdots,U_{n:n}$ a.s. 唯一确定. 为方便计, 以 P 表示真分布及涉及的概率计算, 以 $\mathbb{U}_n$ 表示基于上述样本的经验分布及涉及的概率计算. 记 $\psi_\zeta(u)=I(u\leqslant\zeta)-p$, $\hat{\zeta}_n=U_{r(n):n}$, 则 $\zeta_0=p$ 为 $P\psi_\zeta=\zeta-p$ 的唯一零点, 且 $\mathbb{U}_n\psi_{\hat{\zeta}_n}=r(n)/n-p=o_p(n^{-1/2})$. 容易验证, 定理 5.19 中的所有条件都满足, 因而 (5.34) 成立:

$$\sqrt{n}(U_{r(n):n}-p)=-\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\big[I(U_i\leqslant p)-p\big]+o_p(1).$$

应用 Delta 方法和反函数的微分规则, 并注意到 $U_i\leqslant p$ 当且仅当 $F^{-1}(U_i)\leqslant F^{-1}(p)$, 即得

$$\sqrt{n}\big[F^{-1}(U_{r(n):n})-F^{-1}(p)\big]=\frac{\sqrt{n}(U_{r(n):n}-p)}{f\big(F^{-1}(p)\big)}+o_p(1)$$
$$=-\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\frac{I\big(F^{-1}(U_i)\leqslant F^{-1}(p)\big)-p}{f\big(F^{-1}(p)\big)}+o_p(1),$$

只要令 $X_i=F^{-1}(U_i)$ 和 $X_{r(n):n}=F^{-1}(U_{r(n):n})$, 即得所要的结论. ■

在 (5.77) 中, 右边和号下的是 n 个 i.i.d. 变量的和, $\sqrt{n}(X_{r(n):n}-\zeta_0)$ 表示成了这样一个线性项和一个余项 R_n 的和, 其中余项 $R_n=o_p(1)$. 这种表示称为 $X_{r(n):n}$ 的**弱线性表示**. 样本分位数也有同样的弱线性表示. 有了这样的表示, 我们马上就可以建立有关次序统计量或样本分位数的渐近分布. 前面在推导 M 估计和极大似然估计的渐近分布时, 也是首先导出了有关估计的类似表达式. 以后我们也常常会这样做. 之所以称这类表示为 Bahadur 表示, 是因为 Bahadur (1966) 首先深入研究了样本分位数的**强线性表示**, 在一定的条件下证得, 以概率 1 当 $n\to\infty$ 时, 余项 R_n 具有 $O\big(n^{-1/4}(\log n)^{3/4}\big)$ 的强收敛速度, 并且引发了一系列重要的后续工作, 例如 Kiefer (1967) (1970a) (1970b). 参看 Serfling (1980).

有了上面的 Bahadur 表示, 我们就可以很容易地导出多维正则中心项序列的中心极限定理.

定理 5.46 设 $X_1,\cdots,X_n$ 是取自分布函数为 F 的一维总体 X 的 i.i.d. 样本, $X_{1:n}\leqslant X_{2:n}\leqslant\cdots\leqslant X_{n:n}$ 为其次序统计量. 设 $p_1<\cdots<p_k$ 为 $(0,1)$ 区间内给定的 k 个数, 自然数 $r_1(n)<\cdots<r_k(n)\leqslant n$. 设对每个 $j=1,\cdots,k$, F 在 p_j 分位数 ζ_j 处有导数 $f(\zeta_j)>0$, 当 $n\to\infty$ 时 $r_j(n)/n=p_j+o(n^{-1/2})$. 记

$$\zeta=(\zeta_1,\cdots,\zeta_k)^\tau,\hat{\zeta}_n=(X_{r_1(n):n},\cdots,X_{r_k(n):n})^\tau,$$

则当 $n\to\infty$ 时, 有

$$\sqrt{n}(\hat{\zeta}_n-\zeta)\rightsquigarrow N_k(0,\Sigma),$$

此处 $\Sigma=(\sigma_{ij})$ 为 k 阶对称方阵, $\sigma_{ij}=p_i(1-p_j)/[f(\zeta_i)f(\zeta_j)],i\leqslant j$.

证明 由 Bahadur 表示, $\sqrt{n}(\hat{\zeta}_n-\zeta)$ 与

$$\frac{1}{\sqrt{n}}\sum_{i=1}^{n}Y_i:=\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\Big(\frac{p_1-I(X_i\leqslant\zeta_1)}{f(\zeta_1)},\cdots,\frac{p_k-I(X_i\leqslant\zeta_k)}{f(\zeta_k)}\Big)^\tau$$

有相同的极限分布, 由 i.i.d. 情况下的多维中心极限定理即得所要的结论. ■

注 5.6 定理 5.46 有两个特例值得注意:

(1) $k=1$, 设 $0<p<1$, F 在其 p 分位数 ζ 处有导数 $f(\zeta)>0$, 且 $r(n)/n=p+o(n^{-1/2})$, 则当 $n\to\infty$ 时有

$$\sqrt{n}(X_{r(n):n}-\zeta)\rightsquigarrow N\big(0,p(1-p)/f^2(\zeta)\big),$$

若取 $p=1/2$, 则得 $X_{r(n):n}$ 或样本中位数的极限分布

$$\sqrt{n}(X_{r(n):n}-\zeta)\rightsquigarrow N\big(0,1/[4f^2(\zeta)]\big).$$

(2) $k=2$, 在定理条件下, 有

$$\sqrt{n}\big[(X_{r_2(n):n}\pm X_{r_1(n):n})-(\zeta_2\pm\zeta_1)\big]\rightsquigarrow N(0,\sigma^2),$$

其中

$$\sigma^2 = \frac{p_2(1-p_2)}{f^2(\zeta_2)} \pm \frac{2p_1(1-p_2)}{f(\zeta_1)f(\zeta_2)} + \frac{p_1(1-p_1)}{f^2(\zeta_1)}.$$

类似的结果可用于总体分位数的大样本统计推断中.

2. 极值分布及其吸引场

仍设 $X_1, \cdots, X_n$ 为从一维总体 X 中抽取的 i.i.d. 样本, X 的分布函数为 F, 记

$$X_{n:n} = \max\{X_1, \cdots, X_n\}, \quad X_{1:n} := \min\{X_1, \cdots, X_n\},$$

分别称为取自分布 F 的**样本极大值**和**样本极小值**, 统称**样本极值**. 考虑到 $\min_{1\leqslant i\leqslant n} X_i = -\max_{1\leqslant i\leqslant n}(-X_i)$, 因而在讨论极限分布时只要考虑样本极大值情形.

定义 5.2 若对某个分布函数 F, 存在 $a_n > 0$ 和 $b_n \in \mathbf{R}$, 使 $(X_{n:n} - b_n)/a_n$ 依分布收敛于一个非退化的分布函数 G, 即

$$F^n(a_n x + b_n) \xrightarrow{d} G(x), \tag{5.78}$$

则称 G 为一**极大值分布**. F 称为**底分布**, 且称 F 属于 G 的**吸引场**, 记为 $F \in D(G)$. 类似地, 可定义**极小值分布**. 极大值分布与极小值分布统称**极值分布**.

注意 (5.78) 中的收敛在附录 A.2 开头已有定义, 是指对函数 G 的每个连续点 x, 总有

$$F^n(a_n x + b_n) \to G(x), \quad n \to \infty.$$

极值分布理论的基本问题是: (1) 有哪些分布可作为极值分布? (2) 收敛于某种特定类型极值分布的条件是什么? 这些问题经过许多学者几十年的研究, 已大体上得到了圆满的解决. 集大成的工作当推 Gnedenko (1943) 和 de Haan (1970). 下面仅考虑极大值分布的上述问题.

为了叙述方便, 引进下面的概念. 两个非退化分布函数 F_1 和 F_2 称为**是同一型的**, 若存在常数 $a > 0$ 和 $b \in \mathbf{R}$, 使 $F_1(x) \equiv F_2(ax + b)$. 在这样一种等价关系之下, 全体一维分布函数被分成一些互斥的子集, 两个分布同型当且仅当它们属于同一子集. 例如, 全体非退化的正态分布即为同一型的. 我们有如下命题.

命题 5.47 *设 U 和 V 为两个非退化的一维分布, X_n 为一串一维随机变量, $a_n > 0, \alpha_n > 0, b_n \in \mathbf{R}, \beta_n \in \mathbf{R}$, 则*

(a) $(X_n - b_n)/a_n \rightsquigarrow U, \quad (X_n - \beta_n)/\alpha_n \rightsquigarrow V,$

蕴含了

(b) $\alpha_n/a_n \to c > 0, \quad (\beta_n - b_n)/a_n \to d \in \mathbf{R},$

以及

(c) $V(x) \equiv U(cx + d)$.

另一方面, 若条件 (b) 成立, 则 (a) 中的任何一个条件都蕴含另一个, 且 (c) 成立.

证明可参看 Resnick (1987) 中的命题 0.2.

关于极值分布的类型, 我们不加证明地引述下面的定理.

定理 5.48 每个极大值分布 G 属于下列各型中的一个型:

$$G_{1,\alpha}(x)=\begin{cases}0, & x\leqslant 0,\\ \exp\{-x^{-\alpha}\}, & x>0,\end{cases}$$

$$G_{2,\alpha}(x)=\begin{cases}\exp\{-(-x)^{\alpha}\}, & x<0,\\ 1, & x\geqslant 0,\end{cases}$$

$$G_3(x)=\exp\{-\mathrm{e}^{-x}\},\qquad -\infty<x<\infty,$$

其中 $\alpha>0$ 为一形状参数 (不同的 α 相应于不同的型). 以上三大类分布分别称为 Fréchet 分布, Weibull 分布和 Gumbel 分布.

参看 Gnedenko (1943), de Haan (1970) (1976), Weissman (1975), Resnick (1987) 以及 Reiss (1989).

为方便计, 将 G_3 写成 $G_{3,\alpha}$ 的形式, 其中的 α 总是理解为 1. 恒等式

$$G_{1,\alpha}^n(n^{1/\alpha}x)\equiv G_{1,\alpha}(x),$$
$$G_{2,\alpha}^n(n^{-1/\alpha}x)\equiv G_{2,\alpha}(x),$$
$$G_3^n(x+\log n)\equiv G_3(x)$$

表明, $G_{i,\alpha}$ 事实上即为样本极大值的极限分布. 将上述恒等式写成统一的形式, 有

$$G_{i,\alpha}^n(a_nx+b_n)\equiv G_{i,\alpha}(x),$$

其中 a_n,b_n 分别为:

(i) $a_n=n^{1/\alpha},b_n=0,i=1$;

(ii) $a_n=n^{-1/\alpha},b_n=0,i=2$;

(iii) $a_n=1,b_n=\log n,i=3$.

由定义, 如果 (5.78) 成立, 则称 F 属于极大值分布 G 的吸引场. 下面我们讨论 $F\in D(G)$ 的有关条件. 为此, 对于任意的一维分布函数 F, 定义

$$F^{-1}(0):=\alpha(F)=\inf\{x:F(x)>0\},\qquad F^{-1}(1):=\omega(F)=\sup\{x:F(x)<1\}.$$

在上述极值分布中, $G_{21}(x)=\exp\{xI(x<0)\}$ 定义了 "负的" 标准指数分布, 可用作这一研究的出发点. 事实上, 设 $U_{n:n}$ 和 $V_{n:n}$ 分别为取自总体 $U\sim U(0,1)$ 和 $V\sim U(-1,0)$ 的 n 个 i.i.d. 变量的样本极大值, 则

$$P\{n(U_{n:n}-1)\leqslant x\}=(1+x/n)^n,\qquad -n\leqslant x\leqslant 0,$$
$$(1+x/n)^n\to\mathrm{e}^x=G_{21}(x),\qquad x\leqslant 0.$$

即对每个 x, 当 $n\to\infty$ 时, 有

$$P\{n(U_{n:n}-1)\leqslant x\}=P\{nV_{n:n}\leqslant x\}\to G_{21}(x).\tag{5.79}$$

由定理 5.39, $X\overset{d.}{=}F^{-1}(U)$, 且 $X_{n:n}\leqslant x\Longleftrightarrow U_{n:n}\leqslant F(x)$. 由 (5.79), 对每个 x, 当 $n\to\infty$ 时, 有

$$F^n(a_nx+b_n)=P\{X_{n:n}\leqslant a_nx+b_n\}=P\{U_{n:n}\leqslant F(a_nx+b_n)\}$$

$$= P\{nV_{n:n} \leqslant n[F(a_nx+b_n)-1]\}$$
$$= G_{21}\big(n[F(a_nx+b_n)-1]\big) + o(1). \tag{5.80}$$

另一方面, 对 $G \in \{G_{1,\alpha}, G_{2,\alpha}, G_3 : \alpha > 0\}$, 我们可将其写成如下的形式:

$$G(x) = G_{21}\big(\log G(x)\big), \quad x \in \big(\alpha(G), \omega(G)\big). \tag{5.81}$$

由 (5.80) 和 (5.81) 可得: 当 $n \to \infty$ 时,

$$\begin{aligned} &F^n(a_nx+b_n) \to G(x) \text{ 对每个 } x, \\ \Longleftrightarrow\ &n\big(1-F(a_nx+b_n)\big) \to -\log G(x), \text{ 对每个 } x \in \big(\alpha(G), \omega(G)\big). \end{aligned} \tag{5.82}$$

这一熟知的等价性是建立弱收敛于极值分布的充要条件的基本工具之一. 这些条件 [属于Gnedenko (1943) 和 de Haan (1970)] 因其简洁完美的形式被视为经典极值理论的奠基石. 下面我们不加证明地引述这些条件.

定理 5.49 分布 F 属于极值分布 $G_{i,\alpha}$ 的吸引场的充要条件为:

$(1,\alpha)$ $\omega(F) = \infty$, $\lim_{t\to\infty} \bar{F}(tx)/\bar{F}(t) = x^{-\alpha}$, $x > 0$;

$(2,\alpha)$ $\omega(F) < \infty$, $\lim_{t\downarrow 0} \bar{F}\big(\omega(F)+tx\big)/\bar{F}\big(\omega(F)-t\big) = (-x)^{\alpha}$, $x < 0$;

(3) $\lim_{t\uparrow\omega(F)} \bar{F}\big(t+xg(t)\big)/\bar{F}(t) = \mathrm{e}^{-x}$, $-\infty < x < \infty$.

此处, $\bar{F} := 1-F, g(t) = \int_t^{\omega(F)} \bar{F}(y)dy/\bar{F}(t)$. 常数 $a_n > 0$ 和 b_n 可按如下方式选取:

$(1,\alpha)$ $b_n^* = 0$, $a_n^* = F^{-1}(1-1/n)$;

$(2,\alpha)$ $b_n^* = \omega(F)$, $a_n^* = \omega(F) - F^{-1}(1-1/n)$;

(3) $b_n^* = F^{-1}(1-1/n)$, $a_n^* = g(b_n^*)$.

由命题 5.47, 对于常数 a_n 和 b_n 的其他选取, 弱极限为 G 的充要条件为:

$$a_n/a_n^* \to 1 \quad \text{且} \quad (b_n - b_n^*)/a_n \to 0, \quad n \to \infty.$$

参看 Reiss (1989). 证明和其他有关结果可参看 Resnick (1987) 中的命题 0.10, 1.9, 1.11 和 1.13.

例 5.16 假定 $F \in D(G_3)$, 则 $\omega(F)$ 取值有限和无穷都是可能的. $\omega(F) = \infty$ 的一个例子是指数分布 $F(x) = 1-\mathrm{e}^{-x}, x > 0$, 此时若取 $a_n = 1, b_n = \log n$, 则对 $x \in \mathbf{R}$, 当 n 充分大时,

$$F^n(x+\log n) = \big(1-\exp\{-(x+\log n)\}\big)^n = (1-\mathrm{e}^{-x}/n)^n \to \exp\{-\mathrm{e}^{-x}\} = G_3(x),$$

换个记号, 设 $Z_{n:n}$ 为 n 个 i.i.d. 指数 $E(0,1)$ 变量的样本极大值, $Y \sim G_3$, 则有

$$Z_{n:n} - \log n \rightsquigarrow Y. \tag{5.83}$$

Gnedenko (1943) 给出了 $\omega(F) < \infty$ 的一个例子: $F(x) = 0, 1-\exp\{-x/(1-x)\}$, 或 1, 视 $x < 0$, $0 \leqslant x < 1$ 或 $x \geqslant 1$ 而定. 取 $a_n = (1+\log n)^{-2}, b_n = \log n/(1+\log n)$, 则可验证容量为 n 的 i.i.d 样本的极大值 M_n 满足

$$P(M_n \leqslant a_nx+b_n) = F^n(a_nx+b_n) \to G_3(x).$$

为利用前面的结果, 今考虑一种更具启发式的方法如下: 令 $g(x) = x/(1+x) : [0,\infty) \mapsto [0,1)$, 则 $g'(x) = 1/(1+x)^2$, 由中值定理, 在 $\log n$ 和 $Z_{n:n}$ 之间存在 ζ_n, 使

$$M_n - b_n = g(Z_{n:n}) - g(\log n) = (Z_{n:n} - \log n)g'(\zeta_n).$$

由 (5.83), $Z_{n:n}/\log n \xrightarrow{P} 1$, 故 $\zeta_n/\log n \xrightarrow{P} 1$, 进而 $g'(\zeta_n)/g'(\log n) \xrightarrow{P} 1$. 因而

$$(M_n - b_n)/a_n \stackrel{d}{=} [g(Z_{n:n}) - g(\log n)]/g'(\log n) \rightsquigarrow Y.$$

这个例子取自 Resnick (1987).

5.6.3 极值分布的参数估计

在涉及气象、水文、地质等方面的极端灾害, 以及金融、保险等领域的极端事件的研究中, 极值统计都有着重要的应用. 本节只涉及极值分布的参数估计. 为此, 我们将极大值分布写成如下的形式:

$$G_{1,\alpha}^{(\mu,\sigma)}(x) = \begin{cases} 0, & x \leqslant \mu, \\ \exp\big\{-\big(\dfrac{x-\mu}{\sigma}\big)^{-\alpha}\big\}, & x > \mu, \end{cases} \qquad \text{Fréchet}$$

$$G_{2,\alpha}^{(\mu,\sigma)}(x) = \begin{cases} \exp\big\{-\big(-\dfrac{x-\mu}{\sigma}\big)^{\alpha}\big\}, & x < \mu, \\ 1, & x \geqslant \mu, \end{cases} \qquad \text{Weibull}$$

$$G_3^{(\mu,\sigma)}(x) = \exp\Big\{-\exp\big(-\frac{x-\mu}{\sigma}\big)\Big\}, \quad -\infty < x < \infty, \quad \text{Gumbel}$$

其中 μ 和 $\sigma > 0$ 分别为位置和刻度参数, $\alpha > 0$ 为形状参数.

假定 $\xi_1, \xi_2, \cdots$ 是底分布为 F 的一列独立观测数据. 基于这些数据, 极值问题的统计建模常常采用如下的方法: 先将数据分割为等长 m 的分段数据, 按照样本极大值的极限分布理论, 当 m 足够大时, 可以假定相继各段的极大值, 记为 $X_1, \cdots, X_n$, 拟合于某个未知的极值分布. 通常 m 可取为一年的观测数目, 各段的极大值即为年度极大. 知道底分布 F 的有关信息, 有助于界定极大值分布属于哪一大类. 但在一些应用问题中, 通常只记录了在一定时间或空间范围中众多观测值的极端值, 例如一些河段的水位, 只有每年全年最高水位的记录, 并没有细到每天的记录. 这无疑会增加界定极值分布类别的难度.

今设极值分布 G 为具有位置–刻度参数和形状参数的上述三类极值分布之一, G 的上 p 分位数 ζ_p 由 $G(\zeta_p) = 1 - p$ 定义, 此处 $0 < p < 1$. 以通常的术语, ζ_p 称为相应于**回复周期** (return period) $1/p$ 的**回复水平** (return level), 因为在合理的精度范围, 水平 ζ_p 即为平均每 $1/p$ 年期望达到或超过的那个水平. 换句确切的话说, 在任一特定年份, 水平 ζ_p 被超出的概率均为 p. 说得直白一点, 所谓百年一遇的超高水位, 指的就是 $\zeta_{(1/100)}$. ζ_p 这个指标对于多雨地区的桥梁设计或者地震频发地区的防震设计有重要意义. 记 $y_p = -\log(1-p)$, 则

$$\zeta_p = \begin{cases} \mu + \sigma[-\log(1-p)]^{-1/\alpha} = \mu + \sigma {y_p}^{-1/\alpha}, & G = G_{1,\alpha}^{(\mu,\sigma)}, \\ \mu - \sigma[-\log(1-p)]^{1/\alpha} = \mu - \sigma {y_p}^{1/\alpha}, & G = G_{2,\alpha}^{(\mu,\sigma)}, \\ \mu - \sigma\log[-\log(1-p)] = \mu - \sigma\log y_p, & G = G_3^{(\mu,\sigma)}. \end{cases} \tag{5.84}$$

记 $\nu_p := -\log y_p$, 则

$$\zeta_p = \begin{cases} \mu + \sigma \exp\{\nu_p/\alpha\}, & G = G_{1,\alpha}^{(\mu,\sigma)}, \\ \mu - \sigma \exp\{-\nu_p/\alpha\}, & G = G_{2,\alpha}^{(\mu,\sigma)}, \\ \mu + \sigma\nu_p, & G = G_3^{(\mu,\sigma)}. \end{cases} \tag{5.85}$$

图 5.1 即为极值分布的 ζ_p 对 $\nu_p = -\log\log[1/(1-p)]$ 的图像的一个示意图.

图 5.1

由此可见, 对于每个 Fréchet 分布, ζ_p 为 $-\log y_p$ 的凸函数; 对于每个 Weibull 分布, ζ_p 为 $-\log y_p$ 的凹函数; 而对于 Gumbel 分布, 函数曲线则为一条直线, 其斜率为正. 由于采用了这样的坐标变换, 回复水平曲线的上述性状就得到了彰显, 因而可以特别方便地用于模型的表示和核实.

在参数型的极值统计中, 一个首要的问题就是: 如何界定极值数据 $X_1, \cdots, X_n$ 拟合于三大类极值分布的哪一类. 为此, 我们可以利用回复水平曲线的上述性状, 画出样本 $(1-p)$ 分位数 $\hat{\zeta}_p$ 对 $-\log y_p$ 的函数曲线, 根据曲线的走向就可以初步判断数据来自于三大类极值分布的哪一类. 必要时可以使用拟合优度检验.

极值模型中的渐近有效估计

由于位置–刻度和形状参数都未知的 $G_{1,\alpha}^{(\mu,\sigma)}$ 和 $G_{2,\alpha}^{(\mu,\sigma)}$ 分布类不具有共同支撑, 在考虑渐近有效估计问题时, 我们首先考虑 Gumbel 模型.

首先需要计算 Gumbel 模型中的 Fisher 信息矩阵. 为此, 设 $Y \sim G_3(y) = \exp\{-\mathrm{e}^{-y}\}$, 则 Y 的密度函数为 $g_3(y) = \exp\{-y - \mathrm{e}^{-y}\}$. 记 $Z = \exp(-Y)$, 则 Z 服从标准指数分布 $E(0,1)$, 因而 Y 的矩母函数即为

$$M(t) = \mathrm{E}[\exp\{tY\}] = \mathrm{E}\big(Z^{-t}\big) = \int_0^\infty z^{-t}\mathrm{e}^{-z}dz = \Gamma(1-t), \quad t < 1.$$

由此即可求得

$$\mathrm{E}(Y)=dM(t)/dt\big|_{t=0}=\gamma,\quad \mathrm{E}(Y^2)=d^2M(t)/dt^2\big|_{t=0}=\gamma^2+\pi^2/6,\quad \mathrm{Var}(Y)=\pi^2/6,$$

其中 $\gamma=\lim_{n\to\infty}(\sum_{k=1}^{n}1/k-\log n)=0.5772\cdots$ 为 Euler 常数 (参看 Johnson 和 Kotz (1970b)).

今设 $X\sim G_3^{(\mu,\sigma)}$, 密度函数为 $g_3^{(\mu,\sigma)}$. 则

$$\mathrm{E}(X)=\mu+\sigma\gamma,\quad \mathrm{Var}(X)=\sigma^2\pi^2/6.$$

记 $\theta=(\mu,\sigma)^\tau$, 则 X 的 Fisher 信息矩阵可以写成

$$I(\theta)=\int\Big[\frac{\partial}{\partial\theta}\log g_3^\theta(x)\Big]\Big[\frac{\partial}{\partial\theta^\tau}\log g_3^\theta(x)\Big]dG_3^\theta(x),$$

有

$$\frac{\partial}{\partial\mu}\log g_3^\theta(X)\overset{d.}{=}\frac{1}{\sigma}\big(1-\mathrm{e}^{-Y}\big),\quad \frac{\partial}{\partial\sigma}\log g_3^\theta(X)\overset{d.}{=}\frac{1}{\sigma}\big(Y-1-Y\mathrm{e}^{-Y}\big).$$

利用分部积分, 有 $\mathrm{E}\big(Y\mathrm{e}^{-Y}\big)=\mathrm{E}(Y)-1,\mathrm{E}\big(Y\mathrm{e}^{-2Y}\big)=\mathrm{E}\big(Y\mathrm{e}^{-Y}\big)-1$, $\mathrm{E}\big(Y^2\mathrm{e}^{-2Y}\big)=2\mathrm{E}\big(Y^2\mathrm{e}^{-Y}\big)-2\mathrm{E}\big(Y\mathrm{e}^{-Y}\big)$, 加上 $\mathrm{E}(Y)$ 和 $\mathrm{E}(Y^2)$ 的表达式, 可得 $\mathrm{Var}(Y-1-Y\mathrm{e}^{-Y})=\pi^2/6+(1-\gamma)^2$. 由此即得

$$I(\mu,\sigma)=\sigma^{-2}\begin{pmatrix}1 & (\gamma-1)\\ (\gamma-1) & \pi^2/6+(1-\gamma)^2\end{pmatrix}.\tag{5.86}$$

Fisher 信息矩阵的逆矩阵为

$$I(\mu,\sigma)^{-1}=\frac{6\sigma^2}{\pi^2}\begin{pmatrix}\pi^2/6+(1-\gamma)^2 & (1-\gamma)\\ (1-\gamma) & 1\end{pmatrix}.\tag{5.87}$$

今设 $X_1,\cdots,X_n$ 为取自分布族 $\{G_3^{(\mu,\sigma)},\ -\infty<\mu<\infty,\sigma>0\}$ 的 i.i.d. 样本, 我们关心的是如何估计未知参数 (μ,σ). 文献中提出了多种方法, 例如样本分位数法, 最小二乘法, MLE 等. 例如, 可参看陈希孺 (1999, 2009), Reiss (1989), Coles (2007).

样本分位数法: 简记 $G(x):=\exp(-\mathrm{e}^{-(x-\mu)/\sigma})$, 则 $G(\mu)=\mathrm{e}^{-1}:=p_1\approx0.3679$, $G(\mu+\sigma)=\exp(-\mathrm{e}^{-1}):=p_2\approx0.6922$, 即 μ 和 $\mu+\sigma$ 分别为总体的 p_1 和 p_2 分位数, 可分别用样本 p_1 和 p_2 分位数 ξ_{n,p_1} 和 ξ_{n,p_2} 来估计, 由此即得 (μ,σ) 的估计 $(\xi_{n,p_1},\ \xi_{n,p_2}-\xi_{n,p_1})$.

这种方法计算简单, 且由定理 5.46 即得上述估计的渐近正态性, 但效率较低.

最小二乘 (LS) 法: 设 $X_{1:n}\leqslant X_{2:n}\leqslant\cdots\leqslant X_{n:n}$ 和 $U_{1:n}\leqslant U_{2:n}\leqslant\cdots\leqslant U_{n:n}$ 分别为取自 Gumbel 总体和 $U(0,1)$ 的 n 个独立观测值的次序统计量, 则由定理 5.41 和分布函数 G 的连续性, $\big(G(X_{1:n}),\cdots,G(X_{n:n})\big)\overset{d.}{=}\big(U_{1:n},\cdots,U_{n:n}\big)$, 知 $\mathrm{E}\big(G(X_{k:n})\big)=k/(n+1)$. 为了估计 (μ,σ), 我们令

$$\exp\big[-\mathrm{e}^{-(X_{k:n}-\mu)/\sigma}\big]\equiv G(X_{k:n})=k/(n+1),\quad 1\leqslant k\leqslant n,$$

即可列出如下的估计方程:

$$(X_{k:n}-\mu)/\sigma=-\log\big\{-\log\big[k/(n+1)\big]\big\}:=c_{nk},\quad 1\leqslant k\leqslant n.$$

LS 法就是寻求 (μ,σ) 之值 (μ^*,σ^*), 使偏差平方和 $\sum_{k=1}^{n}\big[(X_{k:n}-\mu)/\sigma-c_{nk}\big]^2$ 达到最小, (μ^*,σ^*) 称为 (μ,σ) 的 LS 估计. LS 法是常用的估计方法之一.

MLE: 在众多估计之中, 最具吸引力的要数 MLE. 其对数似然函数为

$$\log L(\mu,\sigma;x_1,\cdots,x_n)=-n\log\sigma-\sum_{i=1}^{n}\mathrm{e}^{-(x_i-\mu)/\sigma}-\sum_{i=1}^{n}(x_i-\mu)/\sigma.$$

似然方程为

$$\sum_{i=1}^{n}\exp\big[-\frac{x_i-\mu}{\sigma}\big]=n,$$

$$\sum_{i=1}^{n}(x_i-\mu)\Big[1-\exp\big(-\frac{x_i-\mu}{\sigma}\big)\Big]=n\sigma.$$

上述方程可改写为

$$\mu=-\sigma\log\Big[n^{-1}\sum_{i=1}^{n}\exp(-x_i/\sigma)\Big]\quad\text{且}\quad g_n(\sigma)=0,\tag{5.88}$$

其中

$$g_n(\sigma):=\sigma-n^{-1}\sum_{i=1}^{n}x_i+\Big[\sum_{i=1}^{n}x_i\exp(-x_i/\sigma)\Big]\Big/\Big[\sum_{i=1}^{n}\exp(-x_i/\sigma)\Big].\tag{5.89}$$

(5.88) 中的第二个方程 $g_n(\sigma)=0$ 可以用 Newton-Raphson 迭代程序求解. 若将第 k 次迭代结果记为 $\sigma^*_{k,n}$, 则迭代关系式为

$$\sigma^*_{k,n}=\sigma^*_{k-1,n}-g_n(\sigma^*_{k-1,n})/g_n'(\sigma^*_{k-1,n}).$$

迭代的初始值可以采用 σ 的收敛速度为 $O_p(1/\sqrt{n})$ 的任一估计, 例如可用上面样本分位数法得出的估计 $\sigma^*_{0,n}=\hat{\xi}_{n,p_2}-\hat{\xi}_{n,p_1}$. 这里我们指出, 一次迭代的结果 $\sigma^*_{1,n}$ 即与 MLE 具有同样的渐近分布, 进一步的迭代可以改善估计的有限样本性质.

不难验证定理 5.11 的条件成立, 因而似然方程的相合解 $(\hat{\mu}_n,\hat{\sigma}_n)$ 是 (μ,σ) 的最优渐近正态估计 (BANE), 即渐近有效估计, 有

$$\sqrt{n}(\hat{\mu}_n-\mu,\hat{\sigma}_n-\sigma)^\tau\rightsquigarrow N\big((0,0)^\tau,I(\mu,\sigma)^{-1}\big),$$

此处 $I(\mu,\sigma)$ 为 Gumbel 位置–刻度族的 Fisher 信息矩阵. 故由前面的计算, 有

$$\sqrt{n}\begin{pmatrix}\hat{\mu}_n-\mu\\ \hat{\sigma}_n-\sigma\end{pmatrix}\rightsquigarrow N\left(\begin{pmatrix}0\\0\end{pmatrix},\frac{6\sigma^2}{\pi^2}\begin{bmatrix}\pi^2/6+(1-\gamma)^2 & 1-\gamma\\ 1-\gamma & 1\end{bmatrix}\right).\tag{5.90}$$

下面我们指出, 当位置参数为 0 时, 在 Fréchet 分布族中刻度–形状参数的估计, 可以通过一个变量替换, 转化为在 Gumbel 分布族中位置–刻度参数的估计. 为此, 设 $X\sim G^{(0,\sigma)}_{1,1/\alpha},\alpha>0,\sigma>0$. 令 $W=\log X$, 则当 $-\infty<w<\infty$ 时,

$$P(W\leqslant w)=P(X\leqslant\mathrm{e}^w)=\exp\{-\exp[-(w-\log\sigma)/\alpha]\}.$$

因此, 在分布族 $\{G^{(0,\sigma)}_{1,1/\alpha}\}$ 中通过样本 $X_1,\cdots,X_n$ 估计 (σ,α) 的问题, 就转化为在 Gumbel 分布族 $\{G_3^{(\log\sigma,\alpha)}\}$ 中通过样本 $W_1,\cdots,W_n$ 估计位置–刻度参数的问题了.

对于极值统计建模以及本节未涉及的估计问题, 有兴趣的读者可参看 Reiss (1989), 以及 Coles (2007). 为了统一处理三大类极值分布的统计推断, 并避免因错误界定极值分布类型而导致的风险, Coles (2007) 提出了包容三大类极值分布的一个统

一模型, 并讨论了在何种条件下存在参数向量的 MLE, 以及其 MLE 何时具有渐近有效性等问题.

由 (5.85), 设 Gumbel 模型中位置–刻度参数的 MLE 为 $(\hat{\mu}_n,\ \hat{\sigma}_n)$, 则回复水平 ζ_p 的 MLE 即为 $\hat{\mu}_n + \hat{\sigma}_n \nu_p = \hat{\mu}_n - \hat{\sigma}_n \log[-\log(1-p)]$, 它也是渐近有效的. 另两类极值模型的回复水平的 MLE 也可以类似地写出.

5.6.4 L-统计量

设 $X_1, \cdots, X_n$ 为取自一维总体 $X \sim F$ 的 i.i.d. 样本, $(X_{1:n}, \cdots, X_{n:n})$ 为其次序统计量. 通常所说的 **L-统计量**, 即为次序统计量的一个形如 $\sum_{i=1}^n c_{in} X_{i:n}$ 的线性组合, 其中系数 c_{in} 是实数的一个给定的三角阵列. 一般我们只需要考虑形如

$$T_n = \frac{1}{n} \sum_{i=1}^{n} J\big(\frac{i}{n+1}\big) X_{i:n} + \sum_{j=1}^{m} a_j X_{[np_j]:n} \tag{5.91}$$

的 L-统计量, 其中 J 为 $(0,1)$ 区间上的一个给定的函数, $0 < p_1 < \cdots < p_m < 1$, 而 $a_1, \cdots, a_m$ 是非零常数.

为了统计应用的需要, 我们对底分布 F 的形式一般没有什么特别的限制, 基于 L-统计量的统计推断属于非参数统计的范畴, 一般都需要求助于极限分布. 为了避开复杂的推导, 我们将不加证明地介绍 Stigler (1974) 关于此类统计量极限分布的有关结果. 参看 van der Vaart (1998) 和 Serfling (1980), 这两本书中还介绍了推导极限分布的其他方法及有关的参考文献.

Stigler (1974) 研究了形如

$$S_n = \frac{1}{n} \sum_{i=1}^{n} J\big(\frac{i}{n+1}\big) X_{i:n} \tag{5.92}$$

的 L-统计量, 利用 Hájek 投影的方法, 建立了若干结果, 此处我们只介绍下面的这个定理.

定理 5.50 假定 J 为 $(0,1)$ 区间上 a.e. F^{-1} 连续的有界函数, $\mathrm{E}(X^2) < \infty$, 且

$$\sigma^2(J,F) := \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} J\big(F(x)\big) J\big(F(y)\big) \big[F(x \wedge y) - F(x)F(y)\big] dxdy > 0.$$

记 $\sigma^2(S_n) := \mathrm{Var}(S_n)$, 则当 $n \to \infty$ 时, $n\sigma^2(S_n) \to \sigma^2(J,F)$, 且

$$\sqrt{n}\big(S_n - \mathrm{E}(S_n)\big) = -\int \mathbb{G}_n(u) J\big(F(u)\big) du + o_p(1), \tag{5.93}$$

此处 $\mathbb{G}_n(u) = \sqrt{n}\big(\mathbb{F}_n(u) - F(u)\big)$ 为基于样本的经验过程, $\mathbb{F}_n$ 为其经验分布.

$$\sqrt{n}\big(S_n - \mathrm{E}(S_n)\big) \rightsquigarrow N\big(0, \sigma^2(J,F)\big).$$

这里我们特别指出, J 为 a.e. F^{-1} 连续, 等价于 $J\big(F(\cdot)\big)$ a.e. L 连续, 其中 L 为 Lebesgue 测度.

读者也可参看 van der Vaart (1998) 的 22.2 节的一个较为简短的证明. 利用 (5.77) 和 (5.93) 中的 Bahadur 表示, 不难得到形如 (5.91) 的 L-统计量的极限分布. 我们将此证明留给读者作为一个练习.

例 5.17 (切尾均值 (trimmed mean)) 给定 $0<\alpha<1/2$. 统计量

$$S_n=\frac{1}{n(1-2\alpha)}\sum_{\alpha<i/(n+1)<1-\alpha}X_{i:n}$$

称为 (次序样本的) α-**切尾均值**. 当总体分布对称时, S_n 可用于估计分布的对称中心. 与 $\bar{X}$ 相比, S_n 的优点在于它不受样本中因过失性误差导致的异常值 (outlier) 的影响, 是一个有实用意义的估计量. 上式中的 $n(1-2\alpha)$ 也可用和号下的项数来代替.

用 ξ_α 记底分布 F 的 α 分位数. 由 Stigler (1973) 的定理, 只要假定底分布 F 的 α 和 $1-\alpha$ 分位数唯一, 不需要对 F 再加任何其他条件, 即可建立 S_n 的渐近正态性. 特别地, 在 F 关于 0 点对称且在 $\xi_{1-\alpha}$ 处连续的情形, 我们有

$$\sqrt{n}S_n\rightsquigarrow N(0,\sigma^2),$$

其中

$$\sigma^2=\Big\{\int_{-\xi_{1-\alpha}}^{\xi_{1-\alpha}}x^2dF(x)+2\alpha\xi_{1-\alpha}^2\Big\}/(1-2\alpha)^2.$$

此处, 我们对这一结论给出一个简单的证明. 易见, 我们只要对

$$\tilde{S}_n=\frac{1}{n(1-2\alpha)}\sum_{i=[\alpha n]+1}^{[(1-\alpha)n]}X_{i:n}$$

证明同一结论, 此处, $[a]$ 表示不超过 a 的最大整数. 为此, 记

$$V_n=\sum_{i=1}^n I(X_i\leqslant\xi_\alpha),\ W_n=\sum_{i=1}^n I(X_i<\xi_{1-\alpha}),$$

$$Z_n=\sum_{i=1}^n I(X_i\geqslant\xi_{1-\alpha}),\ \zeta_{k_n}=\sum_{i=1}^{k_n}X_{i:n},$$

则 $V_n\sim B(n,\alpha),W_n\sim B(n,1-\alpha),Z_n\sim B(n,\alpha)$, 且

$$\begin{aligned}(1-2\alpha)n^{-1/2}\tilde{S}_n&=n^{-1/2}\sum_{i=[\alpha n]+1}^{[(1-\alpha)n]}X_{i:n}\\&=n^{-1/2}\sum_{i=V_n+1}^{W_n}X_{i:n}+n^{-1/2}\big(\zeta_{V_n}-\zeta_{[\alpha n]}\big)+n^{-1/2}\big(\zeta_{[(1-\alpha)n]}-\zeta_{W_n}\big).\end{aligned}$$

易见,

$$\begin{aligned}&|n^{-1/2}\big(\zeta_{V_n}-\zeta_{[\alpha n]}\big)-n^{-1/2}\xi_\alpha(V_n-[\alpha n])|\\\leqslant\ &n^{-1/2}|V_n-[\alpha n]|\cdot(|X_{[\alpha n]:n}-\xi_\alpha|+|X_{V_n:n}-\xi_\alpha|),\end{aligned}$$

当 $n\to\infty$ 时, 上式右方 $\cdot$ 号之前的部分依概率有界, $\cdot$ 号之后的部分 a.s. 趋于 0, 故

$$n^{-1/2}\big(\zeta_{V_n}-\zeta_{[\alpha n]}\big)=n^{-1/2}\xi_\alpha(V_n-\alpha n)+o_p(1),$$

同样有

$$n^{-1/2}\big(\zeta_{[\alpha n]}-\zeta_{W_n}\big)=n^{-1/2}\xi_{1-\alpha}(Z_n-\alpha n)+o_p(1),$$

注意到

$$\sum_{i=V_n+1}^{W_n} X_{i:n} = \sum_{i=1}^{n} X_i I(\xi_\alpha < X_i < \xi_{1-\alpha}),$$

以及 $\xi_\alpha = -\xi_{1-\alpha}$, 即可得到以下的 Bahadur 表示:

$$(1-2\alpha)n^{-1/2}\tilde{S}_n = n^{-1/2}\sum_{i=1}^{n} X_i I(|X_i| < \xi_{1-\alpha}) + n^{-1/2}\xi_{1-\alpha}(Z_n - V_n) + o_p(1).$$

由独立和的中心极限定理立得切尾均值的极限分布. 不仅如此, 将这一表示式与引理 5.45 相结合, 我们还可得到下面的 Winsor 化均值的极限分布.

例 5.18 (Winsor 化均值 (Winsorized mean)) 给定 $0 < \alpha < 1/2$. 统计量

$$T_n = \frac{1}{n}\Big\{ \sum_{\alpha<i/(n+1)<1-\alpha} X_{i:n} + n\alpha(X_{[n\alpha]:n} + X_{[n(1-\alpha)]:n})\Big\}$$

称为 (次序样本的) α-**Winsor 化均值**. 当底分布对称时, 它也可用于估计分布的对称中心. 与切尾均值一样, Winsor 化均值也可以削弱样本中异常值的影响. 但在删失过程中, Winsor 化均值比切尾均值保留了样本中较多的信息. Winsor 化均值是形如 (5.91) 的 L-统计量的一个特例. 参考例 5.17, 在 F 关于 0 点对称且 $f(\xi_{1-\alpha}) := F'(\xi_{1-\alpha}) > 0$ 的情形, 有

$$\sqrt{n}T_n \rightsquigarrow N(0, \sigma^2),$$

其中

$$\sigma^2 = \int_{-\xi_{1-\alpha}}^{\xi_{1-\alpha}} x^2 dF(x) + 2\alpha\left(\xi_{1-\alpha} + \frac{\alpha}{f(\xi_{1-\alpha})}\right)^2.$$

此处我们提醒读者, 条件 $f(\xi_{1-\alpha}) > 0$ 蕴含了 F 的 α 和 $1-\alpha$ 分位数的唯一性.

5.7 从最小二乘谈起

本节可以看成是对回归分析的一个科普性的简介. 顾名思义, 我们将从最简单的线性回归模型, 最常用的最小二乘 (LS) 估计谈起, 涉及线性模型中的 M 估计、分布的随机加权逼近、广义线性模型和单指标模型等的某些方面. 其目的主要是想对有关方面作一些常识性的介绍, 随意所至, 谈不上有什么系统性. 为了说明问题, 本节只限于考虑估计问题而不涉及较为复杂的统计推断, 而侧重点则放在各类估计的构作及其渐近理论上.

5.7.1 线性回归的最小二乘估计

线性模型是最简单的一类回归模型, 其理论研究和实际应用历史悠久, 有关的理论体系也较为成熟. 对这类模型的研究至今仍长盛不衰, 究其原因, 固然因其仍不

失为统计建模的一种选择, 即便在理论上也有其特有的意义. 因为对简单模型的研究往往是研究复杂模型的基础和前导, 为此, 我们首先考虑线性模型

$$Y_i = x_i^\tau \beta + e_i, \quad i = 1, 2, \cdots, n, \tag{5.94}$$

其中, x_i 为已知的 p 维向量, β 为未知的 p 维回归系数向量, $\{e_i\}$ 为 i.i.d. 随机误差, 具有共同的分布函数 F. 此处只限于考虑回归系数的估计问题. 我们首先介绍 LS 估计及其优良性质.

在模型 (5.94) 中, 关于回归系数 β 的估计, 最为大家熟悉的莫过于 LS 估计, 记为 $\hat{\beta}_n$, 它定义为目标函数 $\sum_{i=1}^n (Y_i - x_i^\tau \beta)^2$ 的最小值点, 即

$$\hat{\beta}_n := \operatorname*{argmin}_{\beta} \sum_{i=1}^n (Y_i - x_i^\tau \beta)^2. \tag{5.95}$$

记

$$S_n = \sum_{i=1}^n x_i x_i^\tau.$$

为避免重复叙述, 下面我们总是假定 $S_{n_0} > 0$ 且 $n \geqslant n_0$. 此时 β 的 LS 估计可写成

$$\hat{\beta}_n = S_n^{-1} \sum_{i=1}^n x_i Y_i.$$

相合性和渐近正态性是 LS 估计最基本的渐近性质. 文献中已经证明了下列结果.

定理 5.51 假定在模型 (5.94) 中, $e_1, e_2, \cdots$ i.i.d., $\mathrm{E}(e_1) = 0$, $0 < \mathrm{Var}(e_1) = \sigma^2 < \infty$, 则当 $n \to \infty$ 时,

$$\hat{\beta}_n \xrightarrow{\text{pr.}} \beta \Longleftrightarrow \hat{\beta}_n \to \beta \text{ a.s.} \Longleftrightarrow S_n^{-1} \to 0. \tag{5.96}$$

参看 Lai, Robbins 和 Wei (1979), Drygas (1976), 以及 Chen (1979).

定理 5.52 假定在模型 (5.94) 中, $e_1, e_2, \cdots$ i.i.d., $\mathrm{E}(e_1) = 0, 0 < \mathrm{Var}(e_1) = \sigma^2 < \infty$, 则当 $n \to \infty$ 时,

$$S_n^{1/2}(\hat{\beta}_n - \beta) \xrightarrow{d.} N(0, \sigma^2 I_p), \tag{5.97}$$

当且仅当

$$d_n^2 := \max_{1 \leqslant i \leqslant n} x_i^\tau S_n^{-1} x_i \to 0. \tag{5.98}$$

参看 Huber (1973). 需要指出的是, 条件 (5.98) 相当于独立和情形的 Lindeberg 条件, 且 $d_n^2 \to 0 \Longrightarrow S_n^{-1} \to 0$.

5.7.2 线性模型中的 M-估计

考虑模型 (5.94). 设 $\rho(\cdot)$ 为在 $\mathbf{R} = (-\infty, \infty)$ 上定义的一个非单调的凸函数, 满足

$$\min_u \mathrm{E}[\rho(e_1 + u) - \rho(e_1)] = 0. \tag{5.99}$$

极值方程

$$\sum_{i=1}^{n}\rho(Y_i-x_i^\tau\hat\beta_n)=\min_{\beta}\sum_{i=1}^{n}\rho(Y_i-x_i^\tau\beta) \tag{5.100}$$

的任意一个解 $\hat\beta_n$ 即称为 β 的一个 M-估计.

文献中有时也把 β 的 M-估计定义为估计方程

$$\sum_{i=1}^{n}\psi(Y_i-x_i^\tau\beta)x_i=0$$

的一个解, 此处 ψ 为一个适当的函数. 下面我们只考虑由凸函数定义的 M-估计.

例 5.19 下面是几个重要的例子.

(a) LS 估计相当于 $\rho(u)=u^2$ 或 $\psi(u)=u$ 的情形, 此时与之相匹配的条件 (5.99) 为 $\mathrm{E}(e_1)=0$.

(b) 中位数回归: 取 $\rho(u)=|u|$, (5.99) 归结为: $u=0$ 为函数 $\mathrm{E}(|e_1+u|-|e_1|)$ 的最小值点, 即 0 为误差变量的中位数. 由

$$\sum_{i=1}^{n}|Y_i-x_i^\tau\hat\beta_n|=\min_{\beta}\sum_{i=1}^{n}|Y_i-x_i^\tau\beta| \tag{5.101}$$

定义的 $\hat\beta_n$ 称为 β 的最小绝对偏差估计, 简记为 LAD 估计.

在这种情形, $\rho(u)=|u|$ 在 $u=0$ 处的左、右导数不等. 当用估计方程定义 β 的 M-估计时, 文献中常取 $\psi(u)=\mathrm{sgn}(u)$. 但估计方程

$$\sum_{i=1}^{n}\mathrm{sgn}(Y_i-x_i^\tau\beta)x_i=0$$

常常没有解. 此时可以对方程作弹性的处理. 例如当 $n\to\infty$ 时, 允许估计序列 $\hat\beta_n$ 满足类似于 $||S_n^{-1/2}\sum_{i=1}^{n}\mathrm{sgn}(Y_i-x_i^\tau\beta)x_i||=o_p(1)$ 的条件. 参看陈希孺, 赵林城 (1996) 的 4.2 节.

(c) 分位数回归: 取 $\rho_p(u)=u[p-I(u<0)]$, (5.99) 等价于 0 为误差 e_i 的 p 分位数. 此时 β 的估计由方程

$$\sum_{i=1}^{n}\rho_p(Y_i-x_i^\tau\hat\beta_n)=\min_{\beta}\sum_{i=1}^{n}\rho_p(Y_i-x_i^\tau\beta) \tag{5.102}$$

给出, 即

$$\hat\beta_n:=\operatorname*{argmin}_{\beta}\sum_{i=1}^{n}(Y_i-x_i^\tau\beta)[p-I(Y_i-x_i^\tau\beta<0)].$$

分位数回归模型因其应用广泛而备受关注.

为了发展 M-估计的渐近理论, 本节总是假定 ρ 为 $\mathbf{R}$ 上的一个非单调的凸函数, ψ_- 和 ψ_+ 分别为 ρ 的左、右导函数, 而 ψ 是大小介于两者之间的一个函数.

关于 M-估计的弱相合性, Zhao, Rao 和 Chen (1993) 建立了以下定理.

定理 5.53 *假定在模型 (5.94) 中,*

(A1) $\mathrm{E}[\psi(e_1)]=0$ (*无偏性*).

(A2) 存在正的常数 c 和 Δ, 使得

$$|\mathrm{E}[\psi(e_1+u)]| \geqslant c|u|, \quad \text{只要 } |u| \leqslant \Delta, \tag{5.103}$$

且

$$\mathrm{E}[\psi^2(e_1 \pm \Delta)] < \infty. \tag{5.104}$$

(A3) $S_n^{-1} \to 0$, 当 $n \to \infty$.

则 $\hat{\beta}_n$ 是 β 的相合估计.

作为定理 5.53 的一个推论, 关于 LAD 估计的(弱)相合性, 我们有如下定理.

定理 5.54 设 $\hat{\beta}_n$ 为由 (5.101) 定义的 β 的一个 LAD 估计. 假定 0 为 F 的一个中位数, 且存在正的常数 c 和 Δ, 使得

$$|1-2F(u)| \geqslant c|u|, \quad \text{只要 } |u| \leqslant \Delta. \tag{5.105}$$

则当 (A3) 成立时, $\hat{\beta}_n$ 是 β 的相合估计.

应当强调的是, 定理 5.53 和定理 5.54 的条件, 是与 LS 估计情形相合性的条件相当的, 因而条件是很弱的. 但是定理 5.53 的证明是有些难度的.

Chen, Bai, Zhao 和 Wu (1992) 以及 Zhao (2002) 分别建立了 β 的 LAD 估计和 M-估计的强相合性.

关于 β 的 LAD 估计 $\hat{\beta}_n$ 的渐近正态性, Chen, Bai, Zhao 和 Wu (1990) 建立了下面的定理 (也可参看预印本: Bai, Chen, Wu 和 Zhao (1987)).

定理 5.55 假定在模型 (5.94) 中, $\mathrm{med}(e_1) = 0$, 且以下条件成立:

(1) $F'(u)$ 在 0 的某个邻域中存在, 在 0 点连续, 且 $f(0) := F'(0) > 0$;

(2) $d_n^2 := \max_{1\leqslant i\leqslant n} x_i^\tau S_n^{-1} x_i \to 0$ (Lideberg 条件),

则当 $n \to \infty$ 时,

$$2f(0)S_n^{1/2}(\hat{\beta}_n - \beta) \xrightarrow{d.} N(0, I_p). \tag{5.106}$$

这一结果后来也为 Pollard (1991) 得到.

陈希孺, 赵林城 (1996) 指出, 定理 5.55 中的条件 (1) 可减弱为以下条件:

(1*) 导数 $f(0) = F'(0)$ 存在且 $f(0) > 0$.

参看 Zhao (2000).

关于 β 的 M-估计 $\hat{\beta}_n$ 的渐近正态性, 有下述结果.

定理 5.56 假定在模型 (5.94) 中, 以下条件成立:

(i) 函数 $G(u) := \mathrm{E}[\psi(e_1+u)]$ 在 $u=0$ 的附近存在、有限, 在 0 点连续, $\mathrm{E}[\psi(e_1)] = G(0) = 0$, 且 $\lambda := G'(0) > 0$;

(ii) $0 < \mathrm{E}[\psi^2(e_1)] := \sigma^2 < \infty$.

且 $\lim_{u\to 0} \mathrm{E}[\psi(e_1+u) - \psi(e_1)]^2 = 0$ (规则性);

(iii) $d_n^2 := \max_{1\leqslant i\leqslant n} x_i^\tau S_n^{-1} x_i \to 0$ (Lideberg 条件),

则当 $n \to \infty$ 时,

$$S_n^{1/2}(\hat{\beta}_n - \beta) \xrightarrow{d.} N(0, \lambda^{-2}\sigma^2 I_p). \tag{5.107}$$

参看 Bai, Rao 和 Wu (1992), 以及 陈希孺, 赵林城 (1996).

下面我们介绍 β 的 M-估计 $\hat{\beta}_n$ 基于随机加权的一种分布逼近. 按照 Rao 和 Zhao (1992) 的建议, 我们可以选取一列非负随机变量 w_i, 使之满足以下条件:

(iv) $w_1, w_2, \cdots$ 独立同分布, $\mathrm{E}(w_1) = \mathrm{Var}(w_1) = 1$, 且序列 $\{w_i\}$ 与 $\{Y_i\}$ 独立.

令 β_n^* 为满足极值条件

$$\sum_{i=1}^{n} w_i \rho(Y_i - x_i^\tau \beta_n^*) = \min_{\beta \in \mathbf{R}^p} \sum_{i=1}^{n} w_i \rho(Y_i - x_i^\tau \beta)$$

的 β 的一个加权 M-估计. 为方便计, 用 P^* 记在 $Y_1, Y_2, \cdots$ 给定的条件下有关的概率计算, 则有如下定理.

定理 5.57 *假定在模型 (5.94) 中, 上述条件*(i)—(iv) *成立, 且存在一个 $\delta > 0$ 使 $E|e_1|^{2+\delta} < \infty$, 则对于 $S_n^{1/2}(\beta_n^* - \hat{\beta}_n)$ 和 $S_n^{1/2}(\hat{\beta}_n - \beta)$ 之间的多元* Kolmogorov-Smirnov *距离, 有*

$$\sup_{u \in \mathbf{R}^p} |P^*\{S_n^{1/2}(\beta_n^* - \hat{\beta}_n) \leqslant u\} - P\{S_n^{1/2}(\hat{\beta}_n - \beta) \leqslant u\}| \xrightarrow{pr.} 0,$$

式中 p-维向量的不等式即逐个坐标的不等式.

因此, 在观测值给定的条件下, 我们有理由用 $S_n^{1/2}(\beta_n^* - \hat{\beta}_n)$ 的条件分布逼近 $S_n^{1/2}(\hat{\beta}_n - \beta)$ 的分布. 由于条件概率 P^* 是在样本序列给定的条件下计算的, 因此这种分布逼近是可以通过计算机实现的. 至于上述收敛的 a.s. 版本, 可参看 Wu 和 Zhao (1999).

在某些场合, 我们会遇到模型 (5.94) 中的自变量 X_i 为随机变量的情形. 最简单的模型就是假定 $\{(X_i, e_i)\}$ 为 i.i.d. 序列. 在中位数回归的情形, 通常还要假定 e_1 在给定 $X_1 = x$ 时的条件中位数 $\mathrm{med}(e_1|x) = 0$, 且对条件分布函数要加上一些条件, 包括在 0 的导数 $f(0|x) > 0$.

5.7.3 广义线性模型

广义线性模型理论是对线性模型的经典理论的重要推广, 在多元统计建模中已经起了重要作用, 它不仅可用于连续型数据的统计建模, 而且特别适用于离散数据、属性数据的统计建模. 参看 Fahmeir 和 Tutz (1994).

此处我们只限于介绍一维广义线性模型. 为此, 设 $(x_1, Y_1), \cdots, (x_n, Y_n)$ 为一组独立数据, x_i 为 p-维已知设计点列, Y_i 服从指数型分布

$$dP_\theta(y_i) = \exp[y_i\theta_i - b(\theta_i)]d\nu(y_i), \tag{5.108}$$

其中, $\theta_i \in \Theta$, 后者为自然参数空间, $b(\cdot)$ 为已知函数, ν 为一 σ-有限测度. 当 Y_i 为连续型变量时, ν 取为 Lebesgue 测度; 而当 Y_i 为离散型变量时, ν 取为计数测度.

由 (5.108), Y_i 的均值 $\mu_i = db(\theta_i)/d\theta_i$.

在广义线性模型中, Y_i 的均值 μ_i 由

$$\mu_i = h(x_i^\tau \beta) \tag{5.109}$$

界定, 其中, $\beta \in \mathbf{R}^p$ 为回归参数向量, β_0 为真参数, h 为一已知的严格单调函数, $g := h^{-1}$ 通常称为联系函数. 这个关系式反映了响应变量通过其均值与自变量的联系. 如果(在某些情形)可选取 $h(t) = db(t)/dt$, 则这样一个联系函数称为自然联系函数. 此时若省去下标 i, 则有

$$x^\tau \beta = g(\mu) = g(db(\theta)/d\theta) = \theta. \tag{5.110}$$

例 5.20 设 Y 为仅取 $0,1$ 值的变量, 记 $\pi = P(Y=1)$, 对 $y = 0,1$ 有

$$P(Y=y) = \pi^y (1-\pi)^{1-y} = (1-\pi)\exp\left[y \log \frac{\pi}{1-\pi}\right],$$

令 $\theta = \log[\pi/(1-\pi)]$, 则

$$P(Y=y) = \exp\{y\theta - \log(1+\mathrm{e}^\theta)\},\ y = 0,1,$$

即 $b(t) = \log(1+\mathrm{e}^t)$.

本例可选取自然联系函数, 为此, 只要取

$$h(t) = db(t)/dt = \mathrm{e}^t/(1+\mathrm{e}^t),$$
$$\mathrm{E}_\beta(Y|x) = h(x^\tau\beta) = \exp(x^\tau\beta)/[1+\exp(x^\tau\beta)].$$

这就是常用的 logistic 回归模型.

似然方程

$$\sum_{i=1}^n x_i H(x_i^\tau\beta)\sigma^{-2}(x_i^\tau\beta)[Y_i - h(x_i^\tau\beta)] = 0 \tag{5.111}$$

的解称为 β_0 的 MLE, 此处

$$H(t) = dh(t)/dt,\ \sigma^2(x_i^\tau\beta) = \mathrm{Var}_\beta(Y_i),$$

而 $\sigma^2(\cdot)$ 由$b(\cdot)$ 和$h(\cdot)$ 决定. 在自然联系函数的情形, 上述方程可简化为

$$\sum_{i=1}^n x_i[Y_i - h(x_i^\tau\beta)] = 0. \tag{5.112}$$

在许多情形, Y_i 不一定服从指数型分布, 而函数 $\sigma^2(\cdot)$ 也难以确定, 此时只要 Y_i 的均值函数 (5.109) 是正确界定的, 则我们可以用以下的拟似然方程

$$l_n(\beta) := \sum_{i=1}^n x_i H(x_i^\tau\beta)\lambda(x_i^\tau\beta)[Y_i - h(x_i^\tau\beta)] = 0 \tag{5.113}$$

的根去估计 β_0, 称为极大拟似然估计, 此处 $\lambda(\cdot) > 0$ 是适当选取的一个函数.

上述模型自然可以推广到因变量 Y_i 为多维的情形.

5.7.4 其他回归模型

1. 因变量受限制的分位数回归模型

如前所说, 分位数回归模型是备受关注的一类回归模型, 在许多领域都有重要的应用. 在这类模型的基础上, 产生了一些较为复杂的变种, 其中因变量受限制或有

删失的模型, 早就引起了经济学界的重视, 值得单独提一提. Tobit 模型和二值响应模型就是其中两个典型的例子.

假定在线性模型

$$Y_i = x_i^\tau \beta + e_i, \qquad i = 1, \cdots, n$$

中仅有 $Y_i^+ = Y_i I(Y_i \geqslant 0)$ 和 x_i 是可观测的, 此处 $\{x_i\}$ 是已知的 p-维向量, $\{e_i\}$ 是随机误差序列, β 是未知的 p-维回归系数向量, 即考虑的模型是

$$Y_i^+ = (x_i^\tau \beta + e_i)^+, \qquad i = 1, \cdots, n.$$

这种对因变量加上非负限制的模型称为有删失的回归模型, 或 Tobit 模型. 应当强调的是, 计量经济理论的许多重要进展都与因变量取值范围受限制的模型有关.

对这种模型的基本假定是

(A1) $e_1, e_2, \cdots$ i.i.d., 其共同的分布 F 的中位数 (或 p-分位数) 为 0, 且 F 在 0 有导数 $f(0) > 0$.

(A2) β 的参数空间 B 是 $\mathbf{R}^p$ 中的一个有界开子集, 其闭包记为 $\overline{B}$.

此处, x_i 可以是设计点列, 也可以是随机变量的观测值. 若 x_i 为某一随机变量的观测值, 则上述条件要作相应的改变. 为了叙述方便, 我们只考虑中位数回归的情形.

基于 $\mathrm{med}(Y_i^+) = (x_i^\tau \beta)^+$ 的事实, Powell (1984) 提出了 β 的 LAD 估计 $\hat{\beta}_n$, 它定义为极值问题

$$\hat{\beta}_n = \underset{\beta \in \overline{B}}{\operatorname{argmin}} \sum_{i=1}^{n} \left| Y_i^+ - (x_i^\tau \beta)^+ \right|$$

的一个 Borel-可测解.

易见, $\sum_{i=1}^n |Y_i^+ - (x_i^\tau \beta)^+|$ 不是 β 的凸函数, 因而 $\hat{\beta}_n$ 的渐近性质的建立是有一定难度的. 关于 $\hat{\beta}_n$ 的相合性, 渐近正态性的有关结果, 读者可参看 Powell (1984), Pollard (1990), 以及 Rao 和 Zhao (1993). 关于 β 的线性假设的检验, 检验统计量的构建及其渐近分布, 可参看 Zhao (2004).

在二值响应模型中,

$$Y^* = \mathrm{sgn}(Y) = \begin{cases} 1, & \text{如果 } Y \geqslant 0, \\ -1, & \text{如果 } Y < 0, \end{cases} \tag{5.114}$$

$$\mathrm{med}(Y | X = x) = x^\tau \beta, \tag{5.115}$$

其中 Y 是不可观测的, 仅有 Y^* 和 q 维解释变量 X 是可观测的.

关于这类模型的适定性条件以及 $\beta^* := \beta / ||\beta||$ 的估计问题, 可参 Manski (1975) (1985).

设 $(X_1, Y_1^*), \cdots, (X_n, Y_n^*)$ 为 (X, Y^*) 的 i.i.d. 观测值, Manski 提出了 β^* 的如下的极大计分估计:

$$\hat{\beta}^* = \underset{b \in B}{\operatorname{argmin}} \frac{1}{n} \sum_{i=1}^{n} Y_i^* \mathrm{sgn}(X_i^\tau b), \tag{5.116}$$

此处, $B=\{b\in\mathbf{R}^q:||b||=1\}$. 这一估计的想法可简单地表述为: 将正确预报和不正确预报的频率之差极大化.

关于这两类模型的一个简单的综述, 可参看 Rao 和 Zhao (1995).

2. 非线性回归模型

这种模型一般可表述为:

$$Y_i=h(x_i,\beta)+e_i,\quad i=1,\cdots,n, \tag{5.117}$$

其中 h 为已知的(关于 β 为非线性的)函数.

3. 单指标模型 (Single index model)

单指标模型广泛应用于医学, 计量经济, 工业等诸多领域. 这类模型一般具有如下的形式:

$$Y=g(X^\tau\beta)+e, \tag{5.118}$$

其中 Y 和 X 分别为响应变量和解释变量, g 为一未知的联系函数, β 为一个 q 维的未知参数向量, e 为不可观测的误差变量. 显然, 要使模型是适定的, 需要加上某些条件. 相应的观测数据模型为

$$Y_i=g(X_i^\tau\beta)+e_i,\quad i=1,\cdots,n. \tag{5.119}$$

文献中研究得较多的一类单指标模型具有如下的均值结构:

$$G(x):=E(Y|X=x)=g(x^\tau\beta). \tag{5.120}$$

基于观测数据 $(X_1,Y_1),\cdots,(X_n,Y_n)$, 文献中提出了对于 β 的方向的不同的估计方法. 此处, 我们简要介绍 Powell, Stock 和 Stocker (1989) 提出的一类估计.

这种方法的出发点是基于下面的考虑: 假定 q-维变量 X 具有连续可微的密度函数 $f(x)$, 则

$$\mathrm{E}\big(f(X)\nabla G(X)\big)=\mathrm{E}\big(f(X)dg(v)/dv|_{v=X^\tau\beta}\big)\beta:=\gamma\beta:=\beta^*.$$

在较弱的一些条件之下, β^* 可以写成如下形式:

$$\beta^*=-2\mathrm{E}\big(Y\nabla f(X)\big). \tag{5.121}$$

以 $\hat f(x)$ 记密度 $f(x)$ 的一个核估计, 则可得到 β^* 的一个估计

$$\hat\beta^*=-\frac{2}{n}\sum_{i=1}^n\nabla\hat f(X_i)Y_i.$$

假定对任何 v 有 $K(v)=K(-v)$. 在上式中代入

$$\nabla\hat f(X_i)=\frac{1}{(n-1)h_n^{q+1}}\sum_{j\leqslant n,j\neq i}\nabla K\Big(\frac{X_i-X_j}{h_n}\Big),$$

即可将 $\hat\beta^*$ 写成一个 U-统计量的形式:

$$\hat\beta^*=-\frac{2}{n(n-1)h_n^{q+1}}\sum_{1\leqslant i<j\leqslant n}\nabla K\Big(\frac{X_i-X_j}{h_n}\Big)(Y_i-Y_j). \tag{5.122}$$

已有文献指出, 如果对模型的非参数成分和参数成分同时加以估计, 则对非参数成分的估计还会有助于参数成分的估计, 使估计效果有所改善.

4. 非参数回归模型

设 (X,Y) 为 $(p+1)$-维随机向量, $\mathrm{E}|Y|<\infty$, 称 $m(x)=\mathrm{E}(Y|X=x)$ 为 Y 关于 X 的回归函数. 若对 $m(\cdot)$ 的形式没有限制, 要由样本 $(X_1,Y_1),\cdots,(X_n,Y_n)$ 对 $m(x)$ 作出推断, 则我们面临的即为非参数回归的统计问题.

以上这些模型在文献中都有广泛深入的讨论.

5.8 习题和补充

5.1 设 $X,X_1,X_2,\cdots$ 同分布 (不需要假定独立性), $\mathrm{E}|X|<\infty$. 证明:

$$\max_{1\leqslant i\leqslant n}|X_i|/n\to 0,\ \text{a.s.}$$

5.2 设 $X_1,X_2,\cdots$ i.i.d., $\mathrm{E}(X_1^2)=\infty$. 证明: $S_n^2=(n-1)^{-1}\sum_{i=1}^n(X_i-\bar X_n)^2\to\infty$, a.s. 此处, $\bar X_n=\sum_{i=1}^n X_i/n$.

5.3 设 $X_1,\cdots,X_n$ i.i.d., $0<\mathrm{E}(X_1^4)<\infty$, $\bar X$ 和 S_n^2 如上题所示. 试求 $n\to\infty$ 时, $Y_n:=\sqrt{n}(\log S_n^2-\log\sigma^2)$ 的极限分布.

5.4 设 X_n 服从二项分布 $B(n,p),0<p<1$. 记 $p_n=X_n/n$. 试求 $n\to\infty$ 时, $Y_n:=\sqrt{n}(\arcsin\sqrt{p_n}-\arcsin\sqrt{p})$ 的极限分布.

5.5 设样本 X 服从二项分布 $B(n,p),0<p<1$, $g(p)=pq,q=1-p$.

(1) 求出 $g(p)$ 的 UMVU 估计 $T_n=\delta_n(X)$.

(2) 区别 $g'(p)\neq 0$ 和 $g'(p)=0$ 两种情形, 分别求出 $\sqrt{n}(T_n-pq)$ 和 $n(T_n-pq)$ 的极限分布.

5.6 设 $X_1,X_2,\cdots$ 独立, $X_i\sim B(1,p_i)$. 证明: 若 $\sum_{i=1}^\infty p_i(1-p_i)=\infty$, 则

$$\sum_{i=1}^n(X_i-p_i)\Big/\Big(\sum_{i=1}^n p_i(1-p_i)\Big)^{1/2}\rightsquigarrow N(0,1).$$

5.7 设 $X_1,\cdots,X_n$ 独立, 有相同的均值 μ, 其方差分别为 $\sigma_1^2,\cdots,\sigma_n^2$.

(1) 证明: 若 $\sum_{i=1}^n\sigma_i^2=o(n^2)$, 则 $\bar X_n$ 是 μ 的一个相合估计.

(2) 今设 $X_i=\mu+\sigma_i e_i$, 其中 $e_1,\cdots,e_n$ i.i.d., $\mathrm{E}(e_1)=0,\mathrm{E}(e_1^2)=1$. 证明: 若

$$\max_{1\leqslant i\leqslant n}\sigma_i^2\Big/\sum_{j=1}^n\sigma_j^2\to 0,$$

则

$$\sqrt{n}(\bar X_n-\mu)/\bar\sigma_n\rightsquigarrow N(0,1),$$

此处 $\bar\sigma_n^2:=n^{-1}\sum_{i=1}^n\sigma_i^2$. 进一步, 若存在 $0<\sigma_0^2<\infty$, 使

$$\bar\sigma_n^2\to\sigma_0^2,$$

则

$$\sqrt{n}(\bar{X}_n - \mu) \rightsquigarrow N(0, \sigma_0^2).$$

(3) 当 $\sigma_i^2 = i^r, 0 < r < 1$ 时, (2) 中前一条件成立而后一不成立. 在此情形有 $n^{(1-r)/2}(\bar{X}_n - \mu) = O_p(1)$.

5.8 设 $X_1, \cdots, X_n$ 独立, 有相同的均值 μ, 其方差分别为 $\sigma_1^2, \cdots, \sigma_n^2$. 考虑 μ 的形如 $T_n \equiv T_n(w) = \sum_{i=1}^n w_{ni} X_i$ 的估计, 此处 $w \equiv w_n = (w_{n1}, \cdots, w_{nn})$ 为非负的权向量, $\sum_{i=1}^n w_{ni} = 1$.

(1) 证明: 所有的估计 $T_n(w)$ 都是无偏的, 且在这些估计中为使有关估计的方差达到最小, 可选取如下最优的权

$$w_{ni} = \sigma_i^{-2} \Big/ \sum_{j=1}^n \sigma_j^{-2},\ i = 1, \cdots, n.$$

将所得估计记为 T_n^*. 计算 $\mathrm{Var}(T_n^*)$, 并证明: 若 $\sum_{j=1}^n \sigma_j^{-2} \to \infty$, 则 T_n^* 是 μ 的一个相合估计.

(2) 今设 $X_i = \mu + \sigma_i e_i$, 其中 $e_1, \cdots, e_n$ i.i.d., $\mathrm{E}(e_1) = 0, \mathrm{E}(e_1^2) = 1$. 证明: 若

$$\max_{1 \leqslant i \leqslant n} \sigma_i^{-2} \Big/ \sum_{j=1}^n \sigma_j^{-2} \to 0,$$

则

$$(\sum_{i=1}^n \sigma_i^{-2})^{1/2}(T_n^* - \mu) \rightsquigarrow N(0, 1).$$

5.9 (1) 为了用 Monte Carlo 方法计算积分

$$I = \int_1^\infty x^{-1} \sin(2\pi x) dx = 0.153\cdots,$$

一种策略是, 采用变量替换 $y = 1/x$, 将积分转换为

$$I = \int_0^1 (1/y) \sin(2\pi/y) dy$$

并用

$$I_n = \frac{1}{n} \sum_{i=1}^n \frac{1}{Y_i} \sin\left(\frac{2\pi}{Y_i}\right)$$

逼近 I, 此处 $Y_1, \cdots, Y_n$ 为取自均匀分布总体 $R(0, 1)$ 的一个 i.i.d. 样本. 问这种方法是否可行? I_n 是否 a.s. 收敛于 I?

(2) 将 (1) 中的积分 I 推广为

$$I_\alpha = \int_1^\infty x^{-\alpha} \sin(2\pi x) dx,$$

此处 $\alpha > 0$. 采用与 (1) 中相同的变量替换, 并对 I_α 构作相应的 Monte Carlo 估计 $I_{n,\alpha}$. 问对什么样的 α, 估计 $I_{n,\alpha}$ 会收敛于 I_α?

(3) 问对什么样的 α, $\sqrt{n}(I_{n,\alpha} - I_\alpha)$ 有一非退化的极限分布? 求此极限分布.

[参看 Ferguson (1996) 第 24 页.]

5.10 (1) 设 $X, X_1, X_2, \cdots$ 为取值 $\mathbf{R}^2$ 的 i.i.d. 变量, X 取值 $(1,0)^\tau, (0,1)^\tau, (0,0)^\tau$ 和 $(-1,-1)^\tau$ 的概率分别为 θ_1, θ_2, θ_3 和 θ_4, 其和为 1. 求 $\sqrt{n}\big(\bar{X} - \mathrm{E}(X)\big)$ 的极限分布.

(2) 设 $X_1, \cdots, X_n$ 为取自参数为 $\lambda > 0$ 的 Poisson 分布的 i.i.d. 样本, $Z_n = (1/n)\sum_{i=1}^n I_{[X_i=1]}$. 求 $\sqrt{n}\big((\bar{X}_n, Z_n)^\tau - (\lambda, \lambda\mathrm{e}^{-\lambda})^\tau\big)$ 的极限分布.

(3) 令 $p_1(\lambda) = \lambda\mathrm{e}^{-\lambda}$. 求在中心化和刻度标准化后, $p_1(\lambda)$ 的 MLE $\hat{p}_1 \equiv p_1(\bar{X}_n)$ 的渐近分布.

(4) 求在中心化和刻度化后 $(Z_n, \hat{p}_1)^\tau$ 的联合渐近分布.

[参看 Ferguson (1996) 第 34 页, 此处稍作改动.]

5.11 设 $X_1, X_2, \cdots$ 为有连续分布的 i.i.d. 随机变量. 称在时刻 k 有一个纪录产生, 如果 $X_k > \max_{i<k} X_i$. 若时刻 k 有一纪录产生, 则令 $R_k = 1$; 否则令其为 0. 则 R_1, $R_2, \cdots$ 是一列独立的 Bernoulli 变量, $P(R_k = 1) = 1 - P(R_k = 0) = 1/k, k = 1, 2, \cdots$. 用 $S_n = \sum_{k=1}^n R_k$ 表示到时刻 n 的纪录数目, 求 $\mathrm{E}(S_n)$ 和 $\sigma_n^2 := \mathrm{Var}(S_n)$, 证明 $(S_n - \mathrm{E}(S_n))/\sigma_n \rightsquigarrow N(0,1)$.

[参看 Ferguson (1996) 第 34 页.]

5.12 设 $X_1, \cdots, X_n$ 为 i.i.d. $N(0,1)$ 变量, 并记 $Y_i = X_i^2, i = 1, \cdots, n$, 则 $\sum_{i=1}^n Y_i \sim \chi_n^2$.

(1) 求 $\sqrt{n}(\bar{Y}_n - 1)$ 的极限分布.

(2) 证明: 对每个 $r > 0$, $\sqrt{n}(\bar{Y}_n^r - 1) \rightsquigarrow N\big(0, V^2(r)\big)$. 给出 $V^2(r)$ 的表达式.

(3) 证明 $\sqrt{n}\big[\bar{Y}_n^{1/3} - (1 - 2/9n)\big]\big/\sqrt{2/9} \rightsquigarrow N(0,1)$, 并将此结果与第 (2) 问的结果加以比较.

[第 (3) 问中的变换称为 χ^2 变量的 Wilson-Hilferty 变换.]

5.13 (**二项逼近**) 设 $X_1, \cdots, X_n$ 独立, $X_k \sim B(1, p_k)$, $S_n = \sum_{k=1}^n X_k$, $T_n \sim B(n, p)$, 则对 $\{1, 2, \cdots, n\}$ 的任一子集 C, $|P(S_n \in C) - P(T_n \in C)| \leqslant \sum_{k=1}^n |p_k - p|$.

5.14 (**Poisson 逼近**) 设 $X_{n1}, \cdots, X_{nn}$ 为独立的 Bernoulli 变量, $X_{nk} \sim B(1, p_{nk})$, 且 T_n 服从参数为 $\sum_{k=1}^n p_{nk}$ 的 Poisson 分布. 用 P_n 记 $S_n = \sum_{k=1}^n X_{nk}$ 的分布, Q_n 记 T_n 的分布 $P(\sum_{k=1}^n p_{nk})$. 证明

$$d_{TV}(P_n, Q_n) \equiv \sup_{A \in \mathcal{B}_1} |P(S_n \in A) - P(T_n \in A)| \leqslant \sum_{k=1}^n p_{nk}^2.$$

注意若对所有的 $k \leqslant n$, 令 $p_{nk} = p_n \to 0$ 且 $np_n \to \lambda > 0$, 则 $\sum_{k=1}^n p_{nk}^2 = np_n^2 = (np_n)^2/n = O(1/n)$.

[选自 Ferguson (1996) 第 18 页. 提示: 在同一个概率空间上构造 S_n 和 T_n 如次: 令 $Y_{nk} \sim P(p_{nk}), k = 1, \cdots, n$ 独立, $Z_{nk} \sim B\big(1, 1 - (1 - p_{nk})\mathrm{e}^{p_{nk}}\big), k = 1, \cdots, n$ 独立, 且 $\{Z_{nk}\}$ 与 $\{Y_{nk}\}$ 独立. 定义

$$X_{nk} = I_{[Y_{nk} \geqslant 1]} + I_{[Y_{nk}=0]} I_{[Z_{nk}=1]}.$$

置 $S_n=\sum_{k=1}^n X_{nk}, T_n=\sum_{k=1}^n Y_{nk}$, 则 $X_{nk}\sim B(1,\ p_{nk}), T_n\sim P(\sum_{k=1}^n p_{nk})$, 且

$$P(Y_{nk}=0,X_{nk}=1)=\mathrm{e}^{-p_{nk}}-(1-p_{nk}),$$
$$P(Y_{nk}\geqslant 1,X_{nk}=0)=0,$$
$$P(Y_{nk}\geqslant 2)=1-\mathrm{e}^{-p_{nk}}-p_{nk}\mathrm{e}^{-p_{nk}}.]$$

*5.15 设 $X,X_1,X_2,\cdots$ i.i.d., $\mathrm{E}(X)=0,\mathrm{E}(X^2)=\sigma^2,0<\sigma\leqslant\infty$. 记 $S_n=\sum_{i=1}^n X_i$. 证明: $\lim_{n\to\infty}\mathrm{E}(|S_n|/\sqrt{n})=2\lim_{n\to\infty}\mathrm{E}(S_n^+/\sqrt{n})=\sqrt{2/\pi}\sigma$.
[选自 Chung (1974) 第 172 页. 提示: 当 $\sigma=\infty$ 时, 借助于 X 的对称化, 只要考虑 X 的分布关于 0 对称情形的序列 $\{X_n\}$. 此时, 若对任意的 $c>0$, 令 $Y_n=X_nI(|X_n|\leqslant c), Z_n=X_nI(|X_n|>c)$, 则 $\mathrm{E}(Z_n|Y_n)=0$ a.s.]

5.16 设 $(U_{1:n},\cdots,U_{n:n})$ 为取自 $U(0,1)$ 的次序样本, $R_n=U_{n:n}-U_{1:n}$, 求 $n\to\infty$ 时 $2n(1-R_n)$ 的极限分布.

5.17 设 $X_1,\cdots,X_n$ 为 X 的 i.i.d. 样本, $X\sim U(\theta,1/\theta),0<\theta<1$. 求 θ 的极大似然估计 $\hat{\theta}_n$, 并证明它是 θ 的强相合估计.

5.18 设 $X_1,\cdots,X_n$ 有密度函数

$$f(x_1,\cdots,x_n)=\int_0^\infty(\sqrt{2\pi}\sigma)^{-n}\exp\Big\{-\frac{1}{2\sigma^2}\sum_{i=1}^n(x_i-\mu)^2\Big\}L(\sigma^2)d\sigma^2,$$

其中 $\mu\in(-\infty,\infty)$ 为未知参数, $L(\cdot)$ 为 $(0,\infty)$ 上的一个密度函数, 满足 $\int_0^\infty uL(u)du<\infty$. 试求 μ 的极大似然估计, 并证明它是无偏的和弱相合的.

5.19 设 $X_1,\cdots,X_n$ i.i.d., $X_1\sim E(0,\theta),\theta>0$. 又设我们观测到的是 $Y_1,\cdots,Y_n$, 其中 $Y_i=X_i\wedge C$, C 为一给定的正数. 计算 θ 的极大似然估计 $\hat{\theta}_n$ 并讨论它的相合性.

5.20 设 $X_1,\cdots,X_n$ i.i.d., 其共同分布为指数分布 $E(\mu,\theta),-\infty<\mu<\infty,\theta>0$. 求 (μ,θ) 的 MLE $(\hat{\mu},\hat{\theta})$.

5.21 设 $X_1,\cdots,X_n$ i.i.d., 其共同分布为 $E(0,\theta),\theta>0$, $X_{(1)}\leqslant\cdots\leqslant X_{(n)}$ 为它的次序统计量. 在定数删失的机制之下,我们只能观测到 $X_{(1)},\cdots,X_{(r)}$, 其中 $r\geqslant 2$ 是事先指定的一个整数. 求 θ 的 MLE $\hat{\theta}$.

5.22 设 $X_1,\cdots,X_n$ i.i.d., 服从共同的指数分布 $E(\mu,\theta),\mu\in\mathbf{R},\theta>0$, $X_{(1)}\leqslant\cdots\leqslant X_{(n)}$ 为其次序统计量. 在定数删失的机制之下, 我们只能观测到 $X_{(1)},\cdots,X_{(r)}$, 其中 $r\geqslant 2$ 是事先指定的一个整数. 求 (μ,θ) 的 MLE $(\hat{\mu},\hat{\theta})$.

5.23 设 $X_1,\cdots,X_n$ 为取自某总体 X 的 i.i.d. 样本, X 的密度函数为

$$f_\theta(x)=\begin{cases}\theta, & 0\leqslant x\leqslant 1,\\ 1-\theta, & 1<x\leqslant 2,\end{cases}$$

其中 $0<\theta<1$ 为未知参数. 求 θ 的矩估计和极大似然估计, 并比较它们的极限分布.

5.24 设 $X_1,\cdots,X_n$ 为取自指数分布的 i.i.d. 样本, $X_i\sim E(0,1/\theta),\theta>0$, 亦即 X_i 的密度函数 $f(x|\theta)=\theta\mathrm{e}^{-\theta x}I(x>0)$. 易知, 总体的均值、方差分别为 $1/\theta$ 和

$\eta = 1/\theta^2$.

(1) 证明 η 的 MLE 是 $(\bar{X}_n)^2$, 其中 $\bar{X}_n = \sum_{i=1}^n X_i/n$. 进而证明这个估计是相合的, 渐近正态和渐近有效的.

(2) η 的另一估计为样本方差 $S_n^2 = \sum_{i=1}^n (X_i - \bar{X}_n)^2/(n-1)$. 证明这一估计也是相合的和渐近正态的. 比较两个不同估计的渐近方差.

5.25 设 $X_1, \cdots, X_n$ 为取自总体分布为

$$f_\theta(x) = 3\theta^3(x+\theta)^{-4} I(x>0)dx,\ \theta > 0$$

的 i.i.d. 样本.

(1) 设 T_n 为 θ 的 MLE. 证明: 当真参数为 θ 时, $\sqrt{n}(T_n - \theta) \rightsquigarrow N(0, 5\theta^2/3)$.

(2) 假定要估计的是 $P_\theta(X_1 > 1) = [\theta/(1+\theta)]^3 := g(\theta)$, 它的 MLE 为 $g(T_n)$. 求 $\sqrt{n}[g(T_n) - g(\theta)]$ 的极限分布.

5.26 设 X 服从在 1 处截断的 Poisson 分布, 即

$$P_\theta(X = x) = \theta^x/[x!(\mathrm{e}^\theta - 1)],\ x = 1, 2, \cdots,\ \ 0 < \theta < \infty.$$

(1) 证明: 对 X 的 i.i.d. 样本 $X_1, \cdots, X_n$, 似然方程有唯一解 $\hat{\theta}_n$, 且此解为相合的 MLE.

(2) 求 $\sqrt{n}(\hat{\theta}_n - \theta)$ 的极限分布.

5.27 设 $Y_1, \cdots, Y_n$ 相互独立, $Y_i \sim N(a + bx_i, \sigma^2)$, 其中 a, b, σ^2 未知, 所有的 x_i 均为已知实数, 满足 $\sum_{i=1}^n x_i = 0, \sum_{i=1}^n x_i^2 = 1$.

(1) 证明假设检验 "$H: b = 0 \leftrightarrow K: b \neq 0$" 的似然比检验的否定域有以下形式:

$$\frac{(\sum_{i=1}^n x_i Y_i)^2}{\sum_{i=1}^n (Y_i - \bar{Y} - x_i \sum_{j=1}^n x_j Y_j)^2} > c.$$

(2) 为保证真实水平为 $\alpha, 0 < \alpha < 1$, 应取怎样的 c?

5.28 设 $X_1, \cdots, X_n$ 为取自某总体 X 的 i.i.d. 样本, X 的密度函数为

$$f_\theta(x) = 2\Big(\frac{x}{\theta} I_{(0,\theta]}(x) + \frac{1-x}{1-\theta} I_{(\theta,1)}(x)\Big),$$

其中 $0 < \theta < 1$ 为未知参数. 找出 θ 的一个估计 $\bar{\theta}_n$, 使 $\sqrt{n}(\bar{\theta}_n - \theta) = O_p(1), (n \to \infty)$.

*5.29 设 $X_1, \cdots, X_n$; $Y_1, \cdots, Y_n$ 相互独立, 其中$X_i \sim E\big(0, 1/(\lambda_i\theta)\big)$, $Y_i \sim E(0, 1/\lambda_i)$, $i = 1, \cdots, n$. 假定要在 0.05 的水平检验假设 "$H_0: \theta = 1 \leftrightarrow H_1: \theta \neq 1$".

(1) 证明 "$H_0 \leftrightarrow H_1$" 的似然比检验当

$$T_n := \sum_{i=1}^n \left(\log \hat{\theta}_n - 2 \log \frac{R_i \hat{\theta}_n + 1}{R_i + 1} \right) \geqslant k$$

时否定 H_0, 其中 $R_i = X_i/Y_i$, $\hat{\theta}_n$ 满足

$$\frac{n}{\hat{\theta}_n} - 2\sum_{i=1}^n \frac{R_i}{R_i \hat{\theta}_n + 1} = 0.$$

(2) 求当 $\theta = 1$ 而 $n \to \infty$ 时统计量 T_n 的极限分布.

[选自 Knight (1999) 第 504 页.]

5.30 设样本空间为 $\{-2,2,-1,1,0\}$, X 的分布如表 5.3 所示.

表 5.3

	-2	2	-1	1	0
H	$\alpha/2$	$\alpha/2$	$1/2-\alpha$	$1/2-\alpha$	α
K	$c\theta$	$c(1-\theta)$	$(1/2-\alpha)\dfrac{1-c}{1-\alpha}$	$(1/2-\alpha)\dfrac{1-c}{1-\alpha}$	$(1-c)\dfrac{\alpha}{1-\alpha}$

此处 α 固定 (即原假设中只有一个分布), $0\leqslant\theta\leqslant 1$, $\alpha/(2-\alpha)<c<\alpha$, c 给定. 抽样一次.

(1) 求 "$H\leftrightarrow K$" 的真实水平为 α 的似然比检验.

(2) 不作试验, 即取检验函数 $\phi(X)\equiv\alpha$. 比较 (1) 和 (2) 中检验的功效函数, 可从中看出什么事实?

(3) 是否存在 "$H\leftrightarrow K$" 的水平为 α 的 UMP 检验?

[该例是低效似然比检验的一个例子.]

5.31 根据离散变量的单个观测值 X, 检验 X 的概率函数的假设 $H\leftrightarrow K$, 其中 H,K 中的分布如表 5.4 所示.

表 5.4

x	1	2	3	4	5	6	7
H	1/12	1/12	1/12	1/4	1/6	1/6	1/6
K	$a/3$	$b/3$	$c/3$	2/3	0	0	0

此处 $a,b,c\geqslant 0, a+b+c=1$ 为未知参数.

(1) 求 "$H\leftrightarrow K$" 的水平为 $\alpha=0.25$ 的似然比检验, 证明这一检验的否定域为 $X\leqslant 3$.

(2) 证明当 $X=4$ 时否定 H 的检验水平是 0.25, 而在对立假设 K 下一致地优于上述似然比检验.

[提示: 以 $p(x|a,b,c)$ 记参数为 (a,b,c) 的对立假设下 X 的概率函数, 以 $p_0(x)$ 记 X 在 H 下概率函数, 则似然比统计量为 $\lambda=\sup_{(a,b,c)\in K}p(X|a,b,c)/p_0(X)$.]

5.32 设 $X_1,\cdots,X_n$ 是 i.i.d. 一维变量, 具有共同的连续的分布函数 F. 设 F_0 为这样的一个特定的分布函数. 为要检验假设 "$H: F=F_0\leftrightarrow K: F\neq F_0$", 考虑如下的 Cramér-von Mises 检验统计量:

$$C_n\equiv\int_{-\infty}^{\infty}n\big(\mathbb{F}_n(x)-F_0(x)\big)^2dF_0(x).$$

(1) 证明:

$$C_n\overset{d}{=}\int_0^1 n\big(\mathbb{U}_n(t)-t\big)^2dt=\int_0^1\mathbb{B}_n(t)^2dt,$$

此处 $\mathbb{U}_n$ 是 n 个 i.i.d. $U(0,1)$ 随机变量的经验分布函数, $\mathbb{B}_n$ 是其经验过程.

(2) 证明当原假设 H 成立时,

$$C_n \rightsquigarrow \int_0^1 \mathbb{B}(t)^2 dt,$$

此处 $\mathbb{B}$ 为一标准的 Brown 桥过程.

[提示: 利用在 $D[0,1]$ 上 $\mathbb{B}_n \rightsquigarrow \mathbb{B}$ 和连续映照定理.]

(3) 假定真分布 $F \neq F_0$. 证明此时有

$$C_n/n \xrightarrow{\text{a.s.}} c \equiv \int_{-\infty}^{\infty} \big(F(x) - F_0(x)\big)^2 dF_0(x) > 0,$$

因而基于 C_n 的检验对所有 $F \neq F_0$ 是相合的.

5.33 (续题 5.32).

(1) 假定 $F = F_n$ 满足

$$\int \big\{\sqrt{n}\big(F_n(x) - F_0(x)\big) - g(x)\big\}^2 dF_0(x) \to 0,$$

即在 $L_2(F_0)$ 中 $\sqrt{n}\big(F_n(x) - F_0(x)\big) \to g(x)$. 试求 C_n 在局部对立假设 F_n 下 (基于 Brown 桥 $\mathbb{B}$ 和 g) 的极限分布.

(2) 设 c 如题 5.32 第 (3) 问所示. 在题 5.32 的架构下, 证明当真分布 $F \neq F_0$ 时

$$\sqrt{n}(C_n/n - c) \rightsquigarrow N(0, V^2),$$

并求出 V^2.

[提示: 利用 $\sqrt{n}(\mathbb{F}_n - F) \stackrel{d}{=} \mathbb{B}_n(F)$, $\mathbb{B}_n \rightsquigarrow \mathbb{B}$ 和连续映照定理.]

5.34 设 x_n 为一列 p-维向量, $S_n = \sum_{i=1}^n x_i x_i^\tau$ 为 p-阶正定方阵. 证明: 当 $n \to \infty$ 时, $S_n^{-1} x_n \to 0$.

5.35 采用上一题的符号. 证明: 当 $n \to \infty$ 时, 若 $d_n^2 := \max_{1 \leqslant i \leqslant n} x_i^\tau S_n^{-1} x_i \to 0$, 则 $S_n^{-1} \to 0$.

[参看定理 5.52.]

*5.36 Wiener 在时刻 0 位于 0 点, 他想与前方的 Poisson 交谈. 此时 Poisson 位于 a 点, $a > 0$. Wiener 和 Poisson 分别按标准 Wiener 过程 $W(t)$ 和 Poisson(λ) 过程 $P(t)$ 运动. 试求最终他们能够相遇的概率.

[提示: 考虑右连续鞅

$$G_u(t) := \frac{\exp\big\{u\big(W(t) - P(t)\big)\big\}}{\mathrm{E}\{\exp[u\big(W(t) - P(t)\big)]\}}, \quad 0 \leqslant t < \infty,$$

以及停时 $\tau = \inf\{s : W(s) - P(s) > a\}$, 则 $\{G_u(0), G_u(\tau \wedge t)\}$ 为鞅.]

*5.37 设 X, Y 独立同分布, 证明 $\mathrm{E}(1/|X - Y|) = \infty$.

附录

A.1 检验函数空间的一个弱紧性定理

为了能在 3.3 节中将 N-P 基本定理推广到更一般的情形, 需要证明在某种弱收敛之下检验函数空间是紧的, 即任一列检验函数都存在收敛的子序列. 极限函数仍为检验函数是自然满足的. 本节的目的就是建立这样一个弱紧性定理.

本节涉及的函数都是可测空间 $(\mathcal{X}, \mathcal{B})$ 上的可测函数. 设 μ 为 $(\mathcal{X}, \mathcal{B})$ 上的一个测度, 仅仅为了统计假设检验的需要, 不妨假定这个测度是 σ-有限的. 设函数 $u, u_n, n \geqslant 1$ 一致有界, 称 u_n 弱收敛于 u, 若对任一 μ-可积函数 f,

$$\int u_n f \, d\mu \to \int uf \, d\mu, \tag{A.1}$$

记为 $u_n \xrightarrow{w} u$.

从所周知, 如果把 a.e. μ 相等的函数视为同一个函数, 则可对每个 μ-可积函数 f, 定义它的 L_1-范数 $||f|| = \int |f| d\mu$, 使 μ-可积函数空间 $L_1(\mu)$ 成为一线性赋范空间, 其中任意两个函数 f, g 之间的距离即为 $\int |f - g| d\mu$.

一个距离空间, 如果存在一个可数稠密子集, 则称这个空间是可分的. μ-可积函数空间 $L_1(\mu)$ 的可分性与 σ-域 $\mathcal{B}$ 的结构有关. 如果一个 σ-域可由可数个集合生成, 则称这个 σ-域为可分的, 或称其有可数的生成元. 在定义中, 这可数个集合不妨假定是互斥的. 顺便指出, 一个可分 σ-域的子 σ-域不一定是可分的. 一个熟知的例子为, 欧氏空间 $(\mathbf{R}, \mathcal{B}_1)$ 的 Borel-域是可分的, 但由 $\mathbf{R}$ 的至多可数子集及其余集全体构成的子 σ-域却不是可分的. 今设在可测空间 $(\mathcal{X}, \mathcal{B})$ 中, $\mathcal{B}$ 由互斥的集合 $B_1, B_2, \cdots$ 生成, 不妨假定这些集合的 μ 测度都是有限的, 则不难证明, 函数空间 $L_1(\mu)$ 是可分的, 它的一个可数稠密子集由所有形如 $\sum_{i=1}^{n} c_{in} I(B_i), n = 1, 2, \cdots$ 的有理系数的简单函数组成.

关于一致有界的函数序列的弱收敛, 有以下的等价命题.

引理 A.1 一致有界的函数序列 u_n 弱收敛于 u, 当且仅当

$$\int_B u_n d\mu \to \int_B u d\mu, \quad n \to \infty, \quad \text{对 } \mu \text{ 测度有限的任一集合 } B. \tag{A.2}$$

证明 (A.1) ⇒ (A.2) 显然. 今证 (A.2) ⇒ (A.1). 设 f 为任给的 μ-可积函数, M 为 $\{|u|, |u_n|, n \geqslant 1\}$ 的一个一致的上界. 对于任给的 $\epsilon > 0$, 存在简单函数 $g = \sum_{i=1}^{k} c_i I(B_i)$, 使 $\int |f - g| d\mu < \epsilon/M$, 且每个 $\mu(B_i) < \infty$. 因而

$$\begin{aligned}\left|\int (u_n - u) f \, d\mu\right| &\leqslant \left|\int u_n (f-g) d\mu\right| + \left|\int u(f-g) d\mu\right| + \left|\int (u_n - u) g d\mu\right| \\ &\leqslant 2\epsilon + \left|\sum_{i=1}^{k} c_i \int_{B_i} (u_n - u) d\mu\right|,\end{aligned}$$

故 $\limsup_{n\to\infty} \left|\int (u_n - u) f \, d\mu\right| \leqslant 2\epsilon$. 由 ϵ 的任意性即得 $\int u_n f \, d\mu \to \int u f \, d\mu$, 即 (A.1) 成立. ■

下面的定理是证明几种最优检验存在性的理论基础.

定理 A.2 设 μ 为一个欧氏空间上的 σ-有限测度, 则满足 $0 \leqslant \phi \leqslant 1$ 的所有可测函数 ϕ 的集合关于弱收敛 (A.1) 是紧的.

下面的证明并不限于欧氏空间, 它也适用于其 σ-域 $\mathcal{B}$ 为可分的可测空间 $(\mathcal{X}, \mathcal{B})$ 的一般情形.

证明 要证的是, 对于任意给定的一列检验函数 ϕ_n, 存在一个子列 $\phi_{n'}$ 和一个 ϕ, 使

$$\lim \int \phi_{n'} f \, d\mu = \int \phi f \, d\mu$$

对所有 μ-可积的 f 成立.

设 μ^* 为等价于 μ 的有限测度, 则 f 为 μ-可积当且仅当 $f^* = f \, d\mu/d\mu* $ 为 μ^*-可积, 且 $\int \phi f^* d\mu^* = \int \phi f \, d\mu$. 不失一般性, 可设 μ 为有限测度.

如前所述, 在 μ-可积函数空间 $L_1(\mu)$ 中存在一个可数稠密子集 $\{f_k\}$. 令

$$\zeta_n(f) = \int \phi_n f \, d\mu,$$

则对于每个 μ-可积函数 f, 序列 $\{\zeta_n(f)\}$ 有界: $|\zeta_n(f)| \leqslant \int |f| d\mu < \infty$. 因 $\{\zeta_n(f_1)\}$ 为有界点列, 故存在一个收敛子列, 记为 $\{\zeta_{1n}(f_1)\}$; 同样, 有界点列 $\{\zeta_{1n}(f_2)\}$ 有一收敛子列, 记为 $\{\zeta_{2n}(f_2)\}$; $\{\zeta_{2n}(f_3)\}$ 又有一收敛子列 $\{\zeta_{3n}(f_3)\}, \cdots\cdots$, 依此类推, 利用对角线法可选出 ζ_n 的一个子列 ζ_{nn}, 使得对每个 k, 当 $n \to \infty$ 时, 点列 $\zeta_{nn}(f_k)$ 都收敛.

由不等式 $\left|\int (\phi_{nn} - \phi_{mm}) f \, d\mu\right| \leqslant \left|\int (\phi_{nn} - \phi_{mm}) f_k d\mu\right| + 2\int |f - f_k| d\mu$ 及 $\{f_k\}$ 稠于 $L_1(\mu)$, 知对每个 μ-可积函数 f, 数列 $\zeta_{nn}(f)$ 为 Cauchy 列, 故存在与 f 有关的实数 $\zeta(f)$, 使

$$\zeta_{nn}(f) = \int \phi_{nn} f \, d\mu \to \zeta(f), \ n \to \infty. \tag{A.3}$$

取 $f = I(B)$, 即集合 B 的示性函数, 并记

$$\zeta^*(B) = \zeta[I(B)],\ B \in \mathcal{B}. \tag{A.4}$$

由 (A.3), (A.4) 不难看出, ζ^* 为 $\mathcal{B}$ 上的有限可加的非负集合函数, 且对任一 $B \in \mathcal{B}$ 总有 $\zeta^*(B) \leqslant \mu(B)$, 由此及 μ 为有限测度, 易见当 $\mathcal{B}$ 中的集合序列 $B_k \downarrow \emptyset$ 时总有$\zeta^*(B_k) \downarrow 0$, 因而 ζ^* 为 $\mathcal{B}$ 上的一个有限测度, 且 $\mu(B) = 0 \Rightarrow \zeta^*(B) = 0$, 即 $\zeta^* \ll \mu$, 故存在可测函数 ϕ, 使

$$\zeta^*(B) = \int_B \phi\, d\mu,\ B \in \mathcal{B},$$

易见, $0 \leqslant \phi \leqslant 1$, a.e. μ. 由已证之结果, 对每个 $B \in \mathcal{B}$, 总有 $\int_B \phi_{nn} d\mu = \zeta_{nn}(I_B) \to \zeta(I_B) = \int_B \phi\, d\mu$, 故由引理 A.1, $\phi_{nn} \xrightarrow{w.} \phi$. ■

A.2 随机变量序列的各种收敛性

为了第五章阅读的需要, 本节将简要介绍随机向量序列的各种收敛性, 特别注重于依分布收敛和依概率收敛, 并不加证明地陈述有关的结果. 读者可参看 van der Vaart (1998) 一书.

首先, 一个 k-维随机向量 $X = (X_1, \cdots, X_k)^\tau$ 的分布函数, 是指

$$F(x) := P(X \leqslant x) := P(X_1 \leqslant x_1, \cdots, X_k \leqslant x_k), \quad x = (x_1, \cdots, x_k)^\tau \in \mathbf{R}^k.$$

设有一列随机向量 X_n 和一个随机向量 X, 其分布函数分别记为 F_n 和 F. 若对函数 F 的每个连续点 x, 总有

$$F_n(x) \to F(x),\ n \to \infty,$$

则称 X_n **依分布收敛**于 X, 这种收敛又称为**弱收敛**或**依律收敛**, 记为 $X_n \rightsquigarrow X$, 或 $X_n \rightsquigarrow F$, 符号 $\rightsquigarrow$ 也可用 $\xrightarrow{L}$ 或 $\xrightarrow{d}$ 来代替. 顾名思义, 依分布收敛只依赖于有关的随机向量导出的分布律, 不涉及这些随机向量所在的概率空间. 因此这种收敛即为概率测度序列 (或相应的分布函数序列) F_n 的弱收敛, 自然也可记为 $F_n \rightsquigarrow F$.

设 $d(x, y)$ 为 $\mathbf{R}^k$ 中任意两点 x, y 之间的距离. 称一列随机向量 X_n **依概率收敛**于 X, 若对任意给定的 $\varepsilon > 0$,

$$P\big(d(X_n, X) > \varepsilon\big) \to 0,\ n \to \infty,$$

记为 $X_n \xrightarrow{P} X$, 或 $X_n \xrightarrow{\text{pr.}} X$. 按照定义, 它等价于 $d(X_n, X) \xrightarrow{P} 0$. 熟知, 依概率收敛强于依分布收敛, 即 $X_n \xrightarrow{P} X$ 蕴含 $X_n \rightsquigarrow X$.

称 X_n **以概率 1 收敛**于 X, 若

$$P\big(\lim_{n\to\infty} d(X_n, X) = 0\big) = 1,$$

记为 $X_n \to X$, a.s. 或 $X_n \xrightarrow{\text{a.s.}} X$. 以概率 1 收敛又称为 **a.s. 收敛**. 周知, a.s. 收敛是比依概率收敛更强的一种收敛模式.

显然, 在依概率收敛和以概率 1 收敛的定义中, X_n 和 X 必须定义在同一个概率空间上.

今后如不加声明, $\mathbf{R}^k$ 中的距离总取为由欧氏模 $||x|| := \left(\sum_{i=1}^k x_i^2\right)^{1/2}$ 决定的距离.

下面的引理是刻画弱收敛的一揽子等价性引理.

引理 A.3 (一揽子引理, Portmanteau Lemma) 以下每个命题都是随机向量序列 X_n 弱收敛于 X 的等价陈述:

(1) 对所有有界连续函数 f, $\mathrm{E}[f(X_n)] \to \mathrm{E}[f(X)]$;

(2) 对满足 Lipschitz 条件的所有有界函数 f, $\mathrm{E}[f(X_n)] \to \mathrm{E}[f(X)]$;

(3) 对所有非负连续函数 f, $\liminf \mathrm{E}[f(X_n)] \geqslant \mathrm{E}[f(X)]$;

(4) 对每个开集 G, $\liminf P(X_n \in G) \geqslant P(X \in G)$;

(5) 对每个闭集 F, $\limsup P(X_n \in F) \leqslant P(X \in F)$;

(6) $P(X_n \in B) \to P(X \in B)$ 对每个满足 $P(X \in \partial B) = 0$ 的 Borel 集 B 成立, 此处 ∂B 为集合 B 的边界.

以下的连续映照引理虽然简单, 但却很有用.

定理 A.4 (连续映照定理) 设 $X, X_1, X_2, \cdots$ 都是 k-维随机向量, $g : \mathbf{R}^k \mapsto \mathbf{R}^m$ 在集合 $C \subset \mathbf{R}^k$ 的每个点处连续, $P(X \in C) = 1$. 则有

(1) $X_n \rightsquigarrow X$ 蕴含 $g(X_n) \rightsquigarrow g(X)$;

(2) $X_n \xrightarrow{P} X$ 蕴含 $g(X_n) \xrightarrow{P} g(X)$;

(3) $X_n \xrightarrow{\text{a.s.}} X$ 蕴含 $g(X_n) \xrightarrow{\text{a.s.}} g(X)$.

现在我们定义跟弱收敛有关的一些重要概念.

设 $\Pi := \{P_t, t \in T\}$ 为 $(\mathbf{R}^k, \mathcal{B}_k)$ 上的一族概率测度. 称 Π 为**胎紧的** (tight), 如果对于任意给定的 $\varepsilon > 0$, 存在 $\mathbf{R}^k$ 的一个紧子集 K, 使得

$$\inf_t P_t(K) \geqslant 1 - \varepsilon.$$

易见, $(\mathbf{R}^k, \mathcal{B}_k)$ 上的概率测度的任一有限族总是胎紧的. 称 Π 为**相对紧的**, 若 Π 中的概率测度的任一序列, 都存在弱收敛的一个子序列, 此处并不要求其极限元属于 Π.

类似地, 称一族随机向量 $\{X_t, t \in T\}$ 为**胎紧的**, 如果对应的概率分布测度族是胎紧的, 即对于任意给定的 $\varepsilon > 0$, 存在一个常数 M, 使得

$$\sup_t P(||X_t|| > M) < \epsilon.$$

称该随机向量族是**相对紧的**, 如果对应的概率分布族是相对紧的. 由 Markov 不等式不难看出, 如果对某个 $r > 0$, 一族随机变量的 r 阶矩有界, 则该随机变量族就是胎紧的. 对于一列随机向量, 胎紧与依概率有界显然是一回事.

对于一般的距离空间 $(S, \mathcal{B}_S)$ 上的随机元, 已无分布函数可言, 有关的随机元和概率测度序列的弱收敛可以依据一揽子等价性引理加以定义. 对于一族概率测度 Π, 我们也可以类似地定义胎紧和相对紧的概念. 关于胎紧和相对紧之间的关系, 有下面重要的定理.

定理 A.5 (Prohorov 定理) (1) 若 Π 是胎紧的, 则它是相对紧的;

(2) 设 S 为可分的完备距离空间, 若 Π 是相对紧的, 则它是胎紧的.

证明可参看 Billingsley (1986) (1999).

由此立得下面的定理.

定理 A.6 $(\mathbf{R}^k, \mathcal{B}_k)$ 上的一族概率测度 Π 是相对紧的, 当且仅当它是胎紧的.

下面介绍经典的 Helley 选择定理. 为此, 我们引进广义分布函数的概念. 称一个一元函数 F 为广义分布函数, 如果它是非减的, 右连续的, 且 $F(-\infty) \geqslant 0$ 而 $F(+\infty) \leqslant 1$. 多维情形可以类似地定义.

定理 A.7 (Helley 选择定理) $\mathbf{R}^k$ 上的任一分布函数序列 F_n 必有一个子序列 F_{i_n}, 在下述意义下收敛于一个广义分布函数 F: 对于 F 的每个连续点 x, $F_{i_n}(x) \to F(x), n \to \infty$.

Helley 选择定理可用于前一定理的直接证明.

将上述三种收敛性之间的关系总结一下, 即得下面的定理.

定理 A.8 设 X_n, X 和 Y_n 均为 k 维随机向量, 则有

(1) $X_n \xrightarrow{\text{a.s.}} X$ 蕴含 $X_n \xrightarrow{P} X$;

(2) $X_n \xrightarrow{P} X$ 蕴含 $X_n \rightsquigarrow X$;

(3) 设 c 为常向量, 则 $X_n \xrightarrow{P} c$ 当且仅当 $X_n \rightsquigarrow c$;

(4) 设 $X_n \rightsquigarrow X$, 且 $d(X_n, Y_n) \xrightarrow{P} 0$, 则 $Y_n \rightsquigarrow X$;

(5) 设 $X_n \rightsquigarrow X$, 且对某一常向量 c, $Y_n \xrightarrow{P} c$, 则 $(X_n, Y_n) \rightsquigarrow (X, c)$;

(6) 若 $X_n \xrightarrow{P} X$, $Y_n \xrightarrow{P} Y$, 则 $(X_n, Y_n) \xrightarrow{P} (X, Y)$;

(7) (Slutsky 定理) 设 $k = 1$, $X_n \rightsquigarrow X$, 且 $Y_n \rightsquigarrow c$, c 为一常数, 则 $X_n + Y_n \rightsquigarrow X + c$, $Y_n X_n \rightsquigarrow cX$, 且 $Y_n^{-1} X_n \rightsquigarrow c^{-1} X$, 只要 $c \neq 0$. 对于多维情形, 也有相应的结果.

(8) (Cramér-Wold device) 设 X_n, X 均为 k 维随机向量, 则 $X_n \rightsquigarrow X$ 当且仅当对任一 $\eta \in \mathbf{R}^k, \eta^\tau X_n \rightsquigarrow \eta^\tau X$. 用此方法, 即可将证明一列随机向量的依分布收敛, 化简为证明一列一维随机变量的依分布收敛.

引理 A.9 若 $\mathbf{R}^k$ 上的分布函数序列 F_n 弱收敛于一个连续的分布函数 F, 则

$$\sup_x |F_n(x) - F(x)| \to 0, \quad n \to \infty.$$

下面引进一些常用的简便记号. 为此, 设 $\{X_n\}$ 和 $\{Y_n\}$ 为两个 k 维随机向量序列. 记 $X_n = O_p(1)$, 若序列 $\{X_n\}$ 依概率有界; 记 $X_n = o_p(1)$, 若序列 $\{X_n\}$ 依概率收敛于 0. 更一般地,

$$Y_n = O_p(X_n) \iff Y_n = X_n R_n \text{ 且 } R_n = O_p(1);$$
$$Y_n = o_p(X_n) \iff Y_n = X_n R_n \text{ 且 } R_n = o_p(1).$$

下面的定理称为 Skorohod 表示定理, 它把一列随机向量的弱收敛表示为另一个概率空间上的一列随机向量的 a.s. 收敛. 因此这一定理又称为 a.s. 表示定理或强表示定理.

定理 A.10 (Skorohod 表示定理) 设一列随机向量 X_n 依分布收敛于一个随机向量 X_0, 则可在一个合适的概率空间上定义一列随机向量 $\{\tilde{X}_n, n = 0, 1, 2, \cdots\}$, 使得对每个 $n \geqslant 0$, $\tilde{X}_n$ 与 X_n 同分布, 且 $\tilde{X}_n \to \tilde{X}_0$, a.s., $n \to \infty$.

为了叙述一列一维随机变量弱收敛和积分 (即期望) 收敛的关系, 需要重温一致可积的定义. 称一列一维随机变量 $X_n, n = 1, 2, \cdots$ 为**一致可积**, 若

$$\lim_{M\to\infty} \sup_n \mathrm{E}\big(|X_n| I(|X_n| \geqslant M)\big) = 0.$$

显然, 序列 X_n 的一致可积性蕴含了序列 $\mathrm{E}|X_n|$ 的有界性. 另一方面, 若对某个 $r > 1$, 序列 $\{\mathrm{E}|X_n|^r, n \geqslant 1\}$ 有界, 则序列 X_n 一致可积.

定理 A.11 设 $X_n \rightsquigarrow X$, 则 $\mathrm{E}|X| \leqslant \liminf_{n\to\infty} \mathrm{E}|X_n|$.

设 $X_n \rightsquigarrow X$, 且随机变量序列 $\{X_n\}$ 一致可积, 则 $\lim_{n\to\infty} \mathrm{E}(X_n) = \mathrm{E}(X)$.

设 $0 \leqslant X_n \rightsquigarrow X$, 且 $\lim_{n\to\infty} \mathrm{E}(X_n) = \mathrm{E}(X)$, 则随机变量序列 $\{X_n\}$ 一致可积.

证明可参看 Billingsley (1986) (1999).

对于一般情形, 有

定理 A.12 假定 $f : \mathbf{R}^k \mapsto \mathbf{R}$ 可测, $X_n \rightsquigarrow X$, X 以概率 1 在 f 的连续点集中取值, 且随机变量序列 $\{f(X_n)\}$ 一致可积, 则 $\lim_{n\to\infty} \mathrm{E}[f(X_n)] = \mathrm{E}[f(X)]$.

A.3 距离空间上的随机元序列的收敛性

为了适应统计渐近理论的发展, 本节将把随机收敛推广到距离空间上的随机元的一般情形, 并引进一些基本概念、常用符号和重要结果. 由于篇幅限制, 我们将仅对部分结果加以证明. 其他证明可参看 van der Vaart (1998).

记号 (S, d) 通常表示一个距离空间, 其中 d 表示相应的度量或距离. 空间 S 上由所有开集生成的 σ-域, 即包含所有开集的最小 σ-域, 称为空间 S 上的 Borel 域. **今后如无特别声明, 在距离空间上引进的 σ-域均为 Borel 域**. 今设 g 是某个概率空间或距离空间到另一距离空间内的一个映照 (或函数), **称 g 是 Borel-可测的**, 如果 g 关于有关的 Borel σ-域是可测的. 如前所述, **定义于概率空间 $(\Omega, \mathcal{A}, P)$ 上的一个 Borel-可测映照 $X : \Omega \mapsto S$ 称为 取值于S 的一个随机元**. 由上述定义容易证明如下引理.

引理 A.13 一个距离空间到另一距离空间的连续映照是 Borel-可测的.

例 A.1 (空间 $l^\infty(T)$) 对于一个给定的集合 T, 在所有有界函数 $z : T \mapsto \mathbf{R}$ 的线性空间中定义范数 (模)

$$||z||_T := \sup\{|z(t)| : t \in T\}.$$

此时, 范数有限的函数 $z : T \mapsto \mathbf{R}$ 的全体构成一个赋范空间, 记为 $l^\infty(T)$, 上述范数称为一致范数. 由此导出的距离即为一致距离. 为要这一空间是可分的, 当且仅当 T 是可数的.

取 T 为广义直线 $[-\infty,\infty]$ 上的区间 $[a,b]$. 用 $C[a,b]$ 记 $[a,b]$ 上实值连续函数全体所成的空间, 用 $D[a,b]$ 记 $[a,b]$ 上右连续且左极限都存在的实值函数全体所成的空间, 赋予的范数为上述一致范数. 则空间 $C[a,b]$ 是可分的, 而空间 $D[a,b]$ 为不可分的. 容易验证, $C[a,b]\subset D[a,b]\subset l^\infty[a,b]$.

取 T 为赋以距离 ρ 的一个全有界距离空间. 用 $UC(T,\rho)$ 记一致连续的函数 $z:T\mapsto\mathbf{R}$ 的全体所成的空间. 因为在全有界集合上的一致连续函数必为有界函数, 故空间 $UC(T,\rho)$ 为 $l^\infty(T)$ 的子空间. 易见, 可将每个空间 $UC(T,\rho)$ 等同于一个空间 $UC(\bar{T},\rho)$. 事实上, 每个全有界 T 空间的完备化 $\bar{T}$ 是紧的, 而每个全有界空间上的连续函数都有唯一的一个到完备化空间上的连续延拓.

与空间 $C[a,b]$ 类似, 空间 $UC(T,\rho)$ 是可分的. 另外, 其 Borel σ-域等同于由所有坐标投影生成的 σ-域. 对给定的 $t\in T$, 投影 $z\mapsto z(t)$ 是连续的, 因而是 Borel-可测的.

例 A.2 (乘积空间) 设 $(\mathcal{X},d)$ 和 $(\mathcal{Y},e)$ 为一对距离空间, 在乘积集合 $\mathcal{X}\times\mathcal{Y}$ 上定义距离

$$\rho\big((x_1,y_1),(x_2,y_2)\big)=d(x_1,x_2)\vee e(y_1,y_2),$$

则 $(\mathcal{X}\times\mathcal{Y},\rho)$ 即为乘积距离空间. 在此空间中, $(x_n,y_n)\to(x,y)$ 当且仅当 $x_n\to x$, $y_n\to y$ 同时成立.

在乘积距离空间中, 两个因子空间的 Borel-域的乘积 σ-域一般不等于乘积距离的 Borel σ-域, 且后者为大, 为要两者相等, 只需两个因子空间是可分的. 由此不难看出, 如果 $X:\Omega\mapsto\mathcal{X}$ 和 $Y:\Omega\mapsto\mathcal{Y}$ 均为 Borel-可测映照, 则映照 $(X,Y):\Omega\mapsto\mathcal{X}\times\mathcal{Y}$ 关于两个因子 Borel-域的乘积 σ-域可测, 但它一般不是 Borel-可测的. 如果两个因子空间都是可分的, 则它是 Borel-可测的.

由于分布函数无法推广到一般的随机元, 我们将借助一揽子引理来定义距离空间上随机元序列的弱收敛. 因此, 称取值于距离空间 (S,d) 的一列随机元 X_n **弱收敛于**一个随机元 X, 如果对每个有界连续函数 $f:S\mapsto\mathbf{R}$, $\mathrm{E}[f(X_n)]\to\mathrm{E}[f(X)]$, 记为 $X_n\rightsquigarrow X$, 或者在必要时记为 $X_n\overset{d}{\rightsquigarrow}X$. 但有时我们会对一些非随机元情形感兴趣, 因而需要将上述定义加以推广, 使之适用于概率空间 $(\Omega,\mathcal{A},P)$ 到距离空间 S 内的任意映照 X_n 的序列.

为此, 对任意的映照 $X:\Omega\mapsto S$, 我们定义其外期望如下:

$$\mathrm{E}^*[f(X)]=\inf\{\mathrm{E}(U)|U\geqslant f(X),U:\Omega\mapsto\mathbf{R}\text{ 可测且 }\mathrm{E}(U)\text{ 存在}\}.$$

$\mathrm{E}(U)$ 存在是指 $\mathrm{E}(U^+),\mathrm{E}(U^-)$ 中至少有一个为有限. 由此可立得外概率 P^* 的定义. 内期望和内概率也可分别定义为 $\mathrm{E}_*f(X):=-\mathrm{E}^*[-f(X)]$ 和 $P_*(B):=1-P^*(\Omega-B)$.

关于外期望和外概率, 我们不加证明地引述如下事实 (参看 van der Vaart 和 Wellner (1996)).

1° 对任意函数 $T:\Omega\mapsto\bar{\mathbf{R}}$, 存在一个可测函数 $T^*:\Omega\mapsto\bar{\mathbf{R}}$, 使得 (i) $T^*\geqslant T$; (ii) 对于任何可测函数 $U:\Omega\mapsto\bar{\mathbf{R}}$, 只要 $U\geqslant T$, a.s. 即有 $T^*\leqslant U$, a.s. 对于这样的

T^*, 只要 $\mathrm{E}(T^*)$ 存在即有 $\mathrm{E}^*(T)=\mathrm{E}(T^*)$. 且若 $\mathrm{E}^*(T)<\infty$ 则 $\mathrm{E}(T^*)$ 存在. 对于内期望, 我们也可以类似地定义 T_*.

2° 对 Ω 的任何子集 B, 有 (i) $P^*(B)=\mathrm{E}^*(I_B), P_*(B)=\mathrm{E}(I_B)_*$; (ii) 存在一个可测集 $B^*\supset B$, 使 $P(B^*)=P^*(B)$; 且对任何一个这样的 B^*, 有 $I_{B^*}=(I_B)^*$; (iii) $(I_B)^*+(I_{\Omega-B})_*=1$.

因此, **称一列映照 $X_n:\Omega\mapsto S$ 弱收敛到一个随机元 X**, 如果对任一有界连续函数 $f:S\mapsto\mathbf{R}$, 有 $\mathrm{E}^*[f(X)]\to\mathrm{E}[f(X)]$, 记为 $X_n\rightsquigarrow X$. 在这个定义中, 要求极限 X 为一随机元等同于要求 X 为 Borel-可测.

下面我们考虑依概率收敛和 a.s. 收敛. 称一列映照 $X_n:\Omega\mapsto S$ 依概率收敛到 X, 如果对于任意给定的 $\epsilon>0$, $P^*\big(d(X_n,X)>\epsilon\big)\to 0$, 记为 $X_n\xrightarrow{P}X$. 称 $X_n:\Omega\mapsto S$ a.s. 收敛到 X, 如果存在一列 (可测的) 随机变量 Δ_n 使得 $d(X_n,X)\leqslant\Delta_n$ 且 $\Delta_n\xrightarrow{\text{a.s.}}0$, 记为 $X_n\xrightarrow{\text{as*}}X$.

这些定义都不需要假定 X_n 的 Borel-可测性. 但 a.s. 收敛的复杂定义却是不可避免的, 因为即使 X_n 和 X 都是 Borel-可测的, 距离 $d(X_n,X)$ 也未必是一个随机变量.

A.2 节中的一揽子引理、连续映照定理和三种收敛性关系的有关论述, 也都适用于现在的定义, 连同其证明都不需要作实质性改变.

引理 A.14 (一揽子引理) 对于任意的映照 $X_n:\Omega\mapsto S$ 和每个取值于 $\mathcal{S}$ 的随机元 X, 以下陈述都是等价陈述:

(1) 对所有有界连续函数 f, $\mathrm{E}^*[f(X_n)]\to\mathrm{E}[f(X)]$;

(2) 对满足 Lipschitz 条件的所有有界函数 f, $\mathrm{E}^*[f(X_n)]\to\mathrm{E}[f(X)]$;

(3) 对每个开集 G, $\liminf P_*(X_n\in G)\geqslant P(X\in G)$;

(4) 对每个闭集 F, $\limsup P^*(X_n\in F)\leqslant P(X\in F)$;

(5) $P^*(X_n\in B)\to P(X\in B)$ 对每个满足 $P(X\in\partial B)=0$ 的 Borel 集 B 成立, 此处 ∂B 为集合 B 的边界.

定理 A.15 对任意的映照 $X_n,Y_n:\Omega\mapsto S$ 和每个取值于 S 的随机元 X,

(1) $X_n\xrightarrow{\text{as*}}X$ 蕴含 $X_n\xrightarrow{P}X$;

(2) $X_n\xrightarrow{P}X$ 当且仅当对 X_n 的每个子序列 $X_{n'}$, 存在进一步的子序列 $X_{n''}\xrightarrow{\text{as*}}X$;

(3) $X_n\xrightarrow{P}X$ 蕴含 $X_n\rightsquigarrow X$;

(4) 设 c 为常向量, 则 $X_n\xrightarrow{P}c$ 当且仅当 $X_n\rightsquigarrow c$;

(5) 设 $X_n\rightsquigarrow X$, 且 $d(X_n,Y_n)\xrightarrow{P}0$, 则 $Y_n\rightsquigarrow X$;

(6) 设 $X_n\rightsquigarrow X$, 且对某一常向量 c, $Y_n\xrightarrow{P}c$, 则 $(X_n,Y_n)\rightsquigarrow(X,c)$;

(7) 若 $X_n\xrightarrow{P}X$, $Y_n\xrightarrow{P}Y$, 则 $(X_n,Y_n)\xrightarrow{P}(X,Y)$.

以下的连续映照定理表明, 在距离空间 S 上 $X_n\rightsquigarrow X$ 当且仅当对每个连续映照 $g:S\mapsto\mathbf{R}$, $g(X_n)\rightsquigarrow g(X)$.

定理 A.16 (连续映照定理) 设 (S,d) 和 (S^*,d^*) 都是赋予了 Borel σ-域的距离空间, $C\subset S$ 且 $g:C\mapsto S^*$ 为一连续映照, $X_n:\Omega\mapsto S$ 为任意的映照, X 为取值于 S 的随机元, 且 $P(X\in C)=1$. 若 $g(X)$ 是取值于 S^* 的随机元, 则有

(1) $X_n\rightsquigarrow X$ 蕴含 $g(X_n)\rightsquigarrow g(X)$;

(2) $X_n\xrightarrow{P}X$ 蕴含 $g(X_n)\xrightarrow{P}g(X)$;

(3) $X_n\xrightarrow{\text{as*}}X$ 蕴含 $g(X_n)\xrightarrow{\text{as*}}g(X)$.

为了推广 Prohorov 定理, 需要引进**胎紧**的概念. 称距离空间上的一个概率测度 P 是**胎紧的**, 如果对任给的 $\epsilon>0$, 存在一个紧集 K, 使 $P(K)>1-\epsilon$. 一个重要的事实是, 设 S 为一完备可分的距离空间, 则在 S 的 Borel σ-域上的每个概率测度都是**胎紧的**. 因此, 在欧氏空间上的每个概率测度也都是**胎紧的**. 称一族概率测度 Π 是**胎紧的**, 如果对任给的 $\epsilon>0$, 存在一个紧集 K, 使 $P(K)>1-\epsilon$ 对每个 $P\in\Pi$ 都成立. 称距离空间上的一个(Borel-可测的)随机元 X 是**胎紧的**, 如果对任给的 $\epsilon>0$, 存在一个紧集 K, 使 $P(X\notin K)<\epsilon$.

称一列任意的映照 $X_n:\Omega\mapsto S$ **是渐近胎紧的** (asymptotically tight), 如果对任给的 $\epsilon>0$, 存在一个紧集 K, 使得

$$\limsup_{n\to\infty}P^*(X_n\notin K^\delta)<\epsilon,\quad\forall\,\delta>0,$$

此处, $K^\delta=\{y:d(y,K)<\delta\}$.

称一个任意的映照序列 X_n 为**渐近可测的**, 如果

$$\mathrm{E}^*[f(X_n)]-\mathrm{E}_*[f(X_n)]\to 0,\quad\text{对所有有界连续函数 } f:S\mapsto\mathbf{R}.$$

此处 E_* 为内期望, 它可以仿照外期望类似地定义. 特别, 若 X_n 为一距离空间上的一列随机元, 即每个 X_n 都是 Borel-可测的, 则 X_n 显然是渐近可测的.

定理 A.17 (Prohorov 定理) 设 $X_n:\Omega\mapsto S$ 是值域为距离空间的一列任意的映照.

(1) 设 $X_n\rightsquigarrow X$, 其中 X 为一胎紧的随机元, 则序列 X_n 是渐近胎紧和渐近可测的;

(2) 设序列 X_n 是渐近胎紧和渐近可测的, 则存在一个子序列 X_{n_j} 和一个胎紧的随机元 X 使得当 $j\to\infty$ 时有 $X_{n_j}\rightsquigarrow X$.

称 $X=\{X_t,t\in T\}$ 为概率空间 $(\Omega,\mathcal{A},P)$ 上的一个**随机过程**, 如果其中的每个 $X_t:\Omega\mapsto\mathbf{R}$ 均为随机变量, 此处 T 为一个任意的指标集. 对每个给定的 $\omega\in\Omega$, 映照 $t\mapsto X_t(\omega)$ 称为一个**样本轨道**. 如果把 X 视为一个随机函数, 则样本轨道即为它的一个现实. 如果每个样本轨道均为一个有界函数, 则 X 可视为一个映照 $X:\Omega\mapsto l^\infty(T)$.

关于过程序列 X_n 的弱收敛, 容易证明以下的引理:

引理 A.18 设 $S_0\subset S$ 是有同一度量 d 的距离空间. 若 X 和每个 X_n 都取值于 S_0, 则 $X_n\rightsquigarrow X$ (S_0,d) 当且仅当 $X_n\rightsquigarrow X$ (S,d).

因此我们将只考虑在 $l^\infty(T)$ 空间中的弱收敛, 这种刻画将自动转化为在 $C[a,b]$

或 $D[a,b]$ 空间中的弱收敛. 为了引进弱收敛的下述刻画性定理, 我们假定对任给的 $\epsilon>0$, 指标集 T 可以分割为有限多个子集 $T_1,\cdots,T_k$, 使得在每个子集上, 样本轨道 $t\mapsto X_{n,t}$ 的变差 (渐近地) 以大概率不超过 ϵ.

定理 A.19 一列任意的映照 $X_n:\Omega\mapsto l^\infty(T)$ 弱收敛于一个胎紧随机元, 当且仅当以下的两个条件成立:

(i) 对任意的 k 和 T 中任意的 $t_1,\cdots,t_k$, $(X_{n,t_1},\cdots,X_{n,t_k})$ 在 $\mathbf{R}^k$ 中弱收敛;

(ii) 对任给的 $\epsilon,\eta>0$, 存在 T 分成有限多个子集 $T_1,\cdots,T_k$ 的一个分割, 使得

$$\limsup_{n\to\infty}P^*\Big(\sup_i\sup_{s,t\in T_i}|X_{n,s}-X_{n,t}|\geqslant\epsilon\Big)\leqslant\eta.$$

引理 A.20 在上一定理的条件下, 在 T 上存在一个度量 ρ, 使得 (T,ρ) 为全有界的, 且可构造序列 X_n 的弱极限, 使其几乎所有的样本轨道在 $UC(T,\rho)$ 中. 进一步, 若弱极限 X 是零均值 Gauss 过程, 则度量 ρ 可取为 $\rho(s,t)=\big[\mathrm{Var}\,(X_s-X_t)\big]^{1/2}$.

A.4 经验过程

A.4.1 经验分布

设 $X_1,\cdots,X_n$ 为取自分布函数为 F 的一维总体的一个随机样本. 当底分布 F 完全未知时, 我们可定义

$$\mathbb{F}_n(t):=\frac{1}{n}\sum_{i=1}^n I(X_i\leqslant t)$$

为样本 $X_1,\cdots,X_n$ 的**经验分布函数**, 并以此作为 F 的一个自然的估计, 而随机函数

$$\mathbb{G}_n:=\sqrt{n}(\mathbb{F}_n-F)$$

则称为**经验过程**. 我们有如下定理.

定理 A.21 (Glivenko-Cantelli) 设 $X_1,X_2,\cdots$ 为分布为 F 的 i.i.d. 随机变量, 则

$$||\mathbb{F}_n-F||_\infty:=\sup_t|\mathbb{F}_n(t)-F(t)|\xrightarrow{\text{a.s.}}0,\quad n\to\infty.$$

根据 Donsker 定理, 当 F 为 $[0,1]$ 区间上的均匀分布时, 其经验过程序列 $\mathbb{G}_n$ 在 $D[0,1]$ 空间弱收敛于一个标准的 Brown 桥 $\mathbb{B}$, 它是一个零均值 Gauss 过程, 协方差函数 $\mathrm{E}(\mathbb{B}_s\mathbb{B}_t)=s\wedge t-st, 0\leqslant s,t\leqslant 1$. 下面是这个定理的一个推广.

定理 A.22 (Donsker) 设 $X_1,X_2,\cdots$ 为分布为 F 的 i.i.d. 随机变量, 则一列经验过程 $\sqrt{n}(\mathbb{F}_n-F)$ 在 $D[-\infty,\infty]$ 空间弱收敛于一个胎紧的零均值随机元 $\mathbb{G}=\mathbb{B}\circ F$, $\mathbb{B}\circ F(t):=\mathbb{B}\big(F(t)\big)$, 协方差函数为 $F(s\wedge t)-F(s)F(t)$.

这个极限随机元 $\mathbb{B}\circ F$ 称为 F-Brown 桥.

为了深入研究经验过程的弱收敛, 我们需要将上述概念和结果作进一步的推广. 为此, 设 P 为可测空间 $(\mathcal{X},\mathcal{A})$ 上的一个概率分布, $X_1,\cdots,X_n$ 为取自分布 P

的 i.i.d. 样本, **经验分布** $\mathbb{P}_n$ 定义为在这 n 个观测值上的离散均匀分布, 即 $\mathbb{P}_n := (1/n)\sum_{i=1}^n \delta_{X_i}$, 此处 δ_x 为退化于 x 的概率分布. 设 $f : \mathcal{X} \mapsto \mathbf{R}$ 为一给定的可测函数, 我们分别用 $\mathbb{P}_n f$ 和 Pf 记 f 在经验测度 $\mathbb{P}_n$ 和概率测度 P 之下的期望:

$$\mathbb{P}_n f := \frac{1}{n}\sum_{i=1}^n f(X_i), \qquad Pf := \int f dP.$$

借助于一个适当的函数类 $\mathcal{F}$, 经验分布 $\mathbb{P}_n$ 可以表达为 $\{\mathbb{P}_n f, f \in \mathcal{F}\}$, 经验过程即可表达为 $\{\mathbb{G}_n f \equiv \sqrt{n}(\mathbb{P}_n f - Pf), f \in \mathcal{F}\}$.

定义 A.1 可测函数 $f : \mathcal{X} \mapsto \mathbf{R}$ 的一个类 $\mathcal{F}$ 称为 P-Glivenko-Cantelli 类, 如果

$$||\mathbb{P}_n f - Pf||_{\mathcal{F}} := \sup_{f\in\mathcal{F}} |\mathbb{P}_n f - Pf| \xrightarrow{\text{as*}} 0.$$

由中心极限定理, 对于给定的任意有限个满足 $Pf_i^2 < \infty$ 的可测函数 $f_i, i = 1, \cdots, k$,

$$(\mathbb{G}_n f_1, \cdots, \mathbb{G}_n f_k) \rightsquigarrow (\mathbb{G}_P f_1, \cdots, \mathbb{G}_P f_k),$$

式中的右边表示一个零均值多维正态分布, 协方差可表达为

$$\mathrm{E}(\mathbb{G}_P f \mathbb{G}_P f^*) = Pff^* - PfPf^*, \quad f, f^* \in \mathcal{F}.$$

定义 A.2 可测函数 $f : \mathcal{X} \mapsto \mathbf{R}^1$ 的一个类 $\mathcal{F}$ 称为 P-Donsker 类, 如果过程序列 $\{\mathbb{G}_n f, f \in \mathcal{F}\}$ 在空间 $l^\infty(\mathcal{F})$ 中依分布收敛于一个胎紧的极限过程. 此时, 极限过程为一零均值的 Gauss 过程, 其协方差函数由上式给出, 称为 P-Brown 桥.

定义中蕴含的一个要求是, 对于每个给定的 n 和 $X_1, \cdots, X_n$ 的每个现实, 样本轨道 $f \mapsto \mathbb{G}_n f$ 是有界的. 满足这一要求的一个充分条件为, 类 $\mathcal{F}$ 具有一个有限且可积的**包络函数** F. 函数 F 称为函数类 $\mathcal{F}$ 的包络, 如果 $|f(x)| \leqslant F(x) < \infty, \forall x, \forall f$. 此处并不要求函数 F 有界.

为方便计, 称向量值函数 $f : x \mapsto \mathbf{R}^k$ 的一个类 $\mathcal{F}$ 为 Glivenko-Cantelli 类, 或 Donsker 类, 如果当 $f = (f_1, \cdots, f_k)$ 跑遍 $\mathcal{F}$ 时, 对每个 i, 由所有函数 $\{f_i : x \mapsto \mathbf{R}\}$ 组成的函数类为 Glivenko-Cantelli 类, 或 Donsker 类.

一类函数 $\mathcal{F} := \{f\}$ 是否为 Glivenko-Cantelli 类或 Donsker 类, 取决于该函数类在某种意义下的"大小". 为此, 我们引进 $\mathcal{F}$ 关于 $L_r(P)$-模

$$||f||_{P,r} = (P|f|^r)^{1/r}$$

的括号熵的定义. $L_r(P)$ 也用来记满足 $||f||_{P,r} < \infty$ 的所有 $f : \mathcal{X} \mapsto \mathbf{R}$ 的集合.

给定函数 l 和 u, 称满足 $l \leqslant f \leqslant u$ 的所有函数 f 的集合为一个括号 $[l, u]$. 称满足 $P(u-l)^r < \epsilon^r$ 的括号 $[l, u]$ 为 $L_r(P)$ 中的一个 ϵ-括号. **括号数** $N_{[]}\big(\epsilon, \mathcal{F}, L_r(P)\big)$ 定义为覆盖 $\mathcal{F}$ 所需要的 ϵ-括号的最小数目 (涉及的所有 l, u 必须在 $L_r(P)$ 中但不必属于 $\mathcal{F}$). 而 $\log N_{[]}\big(\epsilon, \mathcal{F}, L_r(P)\big)$ 则称为**括号熵**.

定理 A.23 (Glivenko-Cantelli) *为要可测函数 $f : \mathcal{X} \mapsto \mathbf{R}$ 的一个类 $\mathcal{F}$ 为一 P-Glivenko-Cantelli 类, 只要对每个给定的 $\epsilon > 0$, $N_{[]}\big(\epsilon, \mathcal{F}, L_1(P)\big) < \infty$.*

对于我们感兴趣的大多数函数类, 当 $\epsilon \downarrow 0$ 时, 括号熵 $\log N_{[]}\big(\epsilon, \mathcal{F}, L_r(P)\big) \uparrow \infty$. 但对于一个 P-Donsker 类而言, 括号熵的增长不能太快. 为了引进下面的定理, 我们

定义括号积分

$$J_{[]}\big(\delta,\mathcal{F},L_2(P)\big):=\int_0^{\delta}\sqrt{\log N_{[]}\big(t,\mathcal{F},L_2(P)\big)}dt.$$

定理 A.24 (Donsker) *为要 $\mathcal{F}$ 为一 P-Donsker 类, 只需 $J_{[]}\big(1,\mathcal{F},L_2(P)\big)<\infty$.*

作为极大不等式 (引理 A.25) 在经验过程理论中的一个应用, 此处将给出这个定理的证明.

证明 令 $\mathcal{G}=\{f-g:f,g\in\mathcal{F}\}$. 给了覆盖 $\mathcal{F}$ 的一组 $(\epsilon/2)$-括号 $[l_i,u_i]$, 只要取所有这些括号的所有上下限之差, 即可构作覆盖 $\mathcal{G}$ 的一组 ϵ-括号 $[l_i-u_j,u_i-l_j]$. 因此 $N_{[]}\big(\epsilon,\mathcal{G},L_2(P)\big)\leqslant\{N_{[]}\big(\epsilon/2,\mathcal{F},L_2(P)\big)\}^2$. 易见, 以 $\mathcal{G}$ 代替 $\mathcal{F}$ 不会改变熵积分的有限性.

对于给定的 $\delta>0$, 选取覆盖 $\mathcal{F}$ 的最小数目的一组 δ-括号, 用来构作 $\mathcal{F}$ 的一个分割 $\mathcal{F}=\bigcup_i\mathcal{F}_i$, 使得每个 $\mathcal{F}_i$ 的直径小于 δ. 由同一 $\mathcal{F}_i$ 中所有函数之差组成的 $\mathcal{G}$ 的子集, 其中每个函数的 $L_2(P)$-模均小于 δ. 因此, 由后面的引理 A.25, 存在一个 $a(\delta)>0$, 使得

$$\begin{aligned}&\mathrm{E}_P^*[\sup_i\sup_{f,g\in\mathcal{F}_i}|\mathbb{G}_n(f-g)|]\\ \lesssim\ & J_{[]}\big(\delta,\mathcal{F},L_2(P)\big)+\sqrt{n}P^*FI\{F>\sqrt{n}a(\delta)\}:=U+V.\end{aligned}$$

显然, 包络 F 可取为 P-平方可积的.

易见, 对固定的 δ, $V\leqslant a(\delta)^{-1}PF^2I\{F>\sqrt{n}a(\delta)\}\to 0, n\to\infty$. 而当 $\delta\to 0$ 时有 $U\to 0$. 由定理 A.19 借助于 Markov 不等式定理即可得证. ■

例 A.3 (分布函数) 设 $\mathcal{F}$ 为形如 $f_t=I_{(-\infty,t]},t\in\mathbf{R}=(-\infty,\infty)$ 的所有一维示性函数所成的类, 则经验过程 $\mathbb{G}_nf_t$ 即为经典的经验过程 $\sqrt{n}(\mathbb{F}_n-F)$. 上面两个定理即为经典的 Glivenko-Cantelli 定理和 Donsker 定理.

为了给出上述示性函数类的括号数的一个阶, 对于给定的 $\epsilon>0$, 考虑 $\mathbf{R}$ 的一个分割 $-\infty=t_0<t_1<\cdots<t_k=\infty$. 可以选取 $k<2/\epsilon^2$, 使得对每个 i, $F(t_i-)-F(t_{i-1})<\epsilon^2$. 考虑形如 $[I_{(-\infty,t_{i-1}]},I_{(-\infty,t_i)}]$ 的括号, 则每个括号的 $L_2(F)$-长度均小于 ϵ, 因而 $N_{[]}(\epsilon,\mathcal{F},L_2(F))\leqslant 2/\epsilon^2$. 这样, 经典 Donsker 定理也就可以由定理 A.24 得出.

例 A.4 (参数族) 设 $\mathcal{F}=\{f_\theta:\theta\in\Theta\}$ 为可测函数的一个参数族, 参数空间 Θ 为 $\mathbf{R}^k$ 的一个有界子集. 假定存在一个可测函数 m 和 $r\geqslant 1$ 满足 $Pm^r<\infty$, 使得对每个 θ_1,θ_2, 有

$$|f_{\theta_1}(x)-f_{\theta_2}(x)|\leqslant m(x)||\theta_1-\theta_2||,$$

则存在仅依赖于 Θ 和 k 的一个常数 K, 使得对每个 $\epsilon\in(0,\mathrm{diam}\Theta)$, 有

$$N_{[]}(\epsilon||m||_{P,r},\mathcal{F},L_r(P))\leqslant K(\mathrm{diam}\Theta/\epsilon)^k.$$

当 $r=2$ 时函数类 $\mathcal{F}$ 即为 Donsker 类.

为了建立这一上界, 对于给定的 $\epsilon>0$, 只要考虑形如 $[f_{\theta_i}-m\epsilon/2,f_{\theta_i}+m\epsilon/2]$ 的括号. 每个括号的 $L_r(P)$ 长度均为 $\epsilon||m||_{P,r}$. 由 Lipschitz 条件, 只要 $||\theta-\theta_i||\leqslant\epsilon/2$,

则 f_θ 即落入上述括号中. 因此我们可以在 Θ 中适当选取一些网格点 θ_i, 使得以这些点为中心, $\epsilon/2$ 为半径的球的全体能够覆盖函数类 $\mathcal{F}$, 而这些括号的个数满足上述不等式.

例 A.5 (函数类的指标集为紧距离空间的情形) 设 $\mathcal{F}=\{f_\theta:\mathcal{X}\mapsto\mathbf{R},\theta\in\Theta\}$ 为一可测函数类, 以 F 为包络函数, 以紧距离空间 Θ 为指标集, $PF<\infty$, 且对每个 x, 映照 $\theta\mapsto f_\theta(x)$ 为连续映照, 则 $\mathcal{F}$ 的 L_1-括号数是有限的, 因而 $\mathcal{F}$ 为 Glivenko-Cantelli 类.

为证此结论, 对参数的每个开球 B, 定义空间 $\mathcal{X}$ 上的函数 $f_B=\inf\{f_\theta:\theta\in B\}$ 和 $f^B=\sup\{f_\theta:\theta\in B\}$, 以及括号 $[f_B,f^B]$. 设 B_n 为以给定点 θ 为中心, 半径下降于 0 的一列开球, 则由连续性假定, 对逐个 x 有 $f^{B_n}(x)-f_{B_n}(x)\downarrow f_\theta(x)-f_\theta(x)=0$, 由包络函数的 P-可积性和控制收敛定理, $f^{B_n}-f_{B_n}$ 在 $L_1(P)$ 中下降收敛于 0. 这样, 给定 $\epsilon>0$, 对每个 θ, 都存在一个以 θ 为中心的开球 B, 使 $[f_B,f^B]$ 为 $L_1(P)$ 中的一个 ϵ-括号. 上述开球的全体构成了紧空间 Θ 的一个开覆盖, 必可从中选出有限的子覆盖, 相应的那些括号即构成了 $\mathcal{F}$ 的一个有限覆盖.

尽管此处可以证明括号数是有限的, 但无法给出括号数的一个上界.

其他例子可参看 van der Vaart(1998).

A.4.2 极大不等式

本节的主要目的是介绍经验过程的极大不等式, 它们在经验过程弱收敛的理论中起着重要的作用. 我们将采用记号 $\lesssim$ 表示不等号 $\leqslant$ 右方可以不计一个绝对常数的因子, 并用 $\mathrm{Log}\ x$ 来记 $1\vee\log x$.

引理 A.25 *设 $\mathcal{F}$ 为 $\mathcal{X}$ 上实值可测函数的任何一个类, F 为其包络函数, 对每个 $f\in\mathcal{F}$, 都有 $Pf^2<\delta^2$, 则有*

$$\mathrm{E}_P^*||\mathbb{G}_n||_{\mathcal{F}}\lesssim J_{[]}(\delta,\mathcal{F},L_2(P))+\sqrt{n}P^*FI\{F>\sqrt{n}a(\delta)\}.$$

此处, $||\mathbb{G}_n||_{\mathcal{F}}:=\sup\{|\mathbb{G}_nf|:f\in\mathcal{F}\}$, $a(\delta)=\delta/[\mathrm{Log}N_{[]}(\delta,\mathcal{F},L_2(P))]^{1/2}$, 而符号 $I\{F>c\}$ 总是表示集合 $\{F>c\}$ 的示性函数.

这个引理的证明比较复杂, 读者可参看 van der Vaart (1998).

推论 A.26 *设 $\mathcal{F}$ 为 $\mathcal{X}$ 上实值可测函数的任何一个类, F 为其包络函数, 则*

$$\mathrm{E}_P^*||\mathbb{G}_n||_{\mathcal{F}}\lesssim J_{[]}(||F||_{P,2},\mathcal{F},L_2(P)).$$

A.4.3 随机函数

使用熵、Glivenko-Cantelli 类和 Donsker 类等概念, 可以简便地陈述在统计程序的渐近分析中需要的 "规则性" 条件. 例如, 在 M-估计的分析中, 经验过程的理论是控制余项的一个有力的工具. 本节对经验过程的研究, 不仅 $\mathcal{F}$ 中的可测函数 $f_{n,\theta}(\cdot)$ 与 n 及参数 θ 有关, 而且需要考虑 $\mathbb{G}_n$ 作用于 $f_{n,\hat{\theta}_n}(\cdot)$ 的情形, 其中 $\hat{\theta}_n$ 为参数 θ 的一个估计. 这样的函数可写成 $\hat{f}_n(\cdot)=\hat{f}_n(\cdot,X^n)$ 的形式, 称为 "随机函数".

一般地, 我们考虑的 "随机函数" 是这样的可测函数 $x \mapsto \hat{f}_n(x,\omega)$, 对每个 x, 作为 ω 的函数, 与观测值 $X^n := (X_1, \cdots, X_n)$ 定义在同一个概率空间 Ω 上. 记号 $\mathbb{P}_n\hat{f}_n$ 和 $P\hat{f}_n$ 即为 ω 固定时作为函数 $x \mapsto \hat{f}_n(x,\omega)$ 的期望的缩写. 为方便计, 称随机函数 $\hat{f}_n$ 取值于函数类 $\mathcal{F}$, 如果对每个 ω, $\hat{f}_n(\cdot,\omega) \in \mathcal{F}$.

今设 $\mathcal{F}$ 为一 P-Glivenko-Cantelli 类, 且 $\mathcal{F}$ 包含随机函数 $\hat{f}_n$ 的每个序列, 则由 $\sup_{f\in\mathcal{F}} |\mathbb{P}_n f - Pf| \to 0$ a.s., 立得 $|\mathbb{P}_n\hat{f}_n - P\hat{f}_n| \to 0$ a.s. 如果还存在 $f_0 \in \mathcal{F}$ 使 $P\hat{f}_n \to Pf_0$ a.s., 则 $\mathbb{P}_n\hat{f}_n \to Pf_0$ a.s.

类似的原则也适用于 $\mathcal{F}$ 为一 P-Donsker 类且包含 $\hat{f}_n$ 的每个序列的情形. 此时过程 $\mathbb{G}_n f, f \in \mathcal{F}$ "一致地" 依分布收敛于一个 P-Brown 桥过程 $\mathbb{G}_P f, f \in \mathcal{F}$. 由引理 A.20, 极限过程的样本轨道关于标准差度量 $\rho(f,g)$ 一致连续. 如果还存在 $f_0 \in \mathcal{F}$ 使 $\rho(\hat{f}_n, f_0) \xrightarrow{P} 0$, 则 $\mathbb{G}_n\hat{f}_n \rightsquigarrow \mathbb{G}_P f_0$, 即我们有如下引理.

引理 A.27 *假定 $\mathcal{F}$ 为可测函数的一个 P-Donsker 类, $\hat{f}_n$ 是取值于 $\mathcal{F}$ 的一列随机函数, 使 $\int \big(\hat{f}_n(x) - f_0(x)\big)^2 dP(x) \xrightarrow{P} 0$, 其中 $f_0 \in L_2(P)$, 则 $\mathbb{G}_n(\hat{f}_n - f_0) \xrightarrow{P} 0$, 因而 $\mathbb{G}_n(\hat{f}_n) \rightsquigarrow \mathbb{G}_P f_0$.*

证明 不妨假定 $f_0 \in \mathcal{F}$. 赋予 $\mathcal{F}$ 以 $L_2(P)$ 距离, 继而赋予 $\Pi := l^\infty(\mathcal{F}) \times \mathcal{F}$ 以乘积距离. 定义 Π 上的一个实值函数 $g(z,f) = z(f) - z(f_0)$. 今设在 Π 中 $(z_n, f_n) \to (z, f)$. 因为 $|g(z_n,f_n) - g(z,f)| \leqslant 2||z_n - z||_{\mathcal{F}} + |z(f_n) - z(f)|$, 故若映照 $z(\cdot)$ 在 f 处连续, 则 g 在点 (z,f) 处连续.

因为 $\mathcal{F}$ 为一 P-Donsker 类, 在 $l^\infty(\mathcal{F})$ 空间中 $\mathbb{G}_n \rightsquigarrow \mathbb{G}_P$. 又由假定, 在距离空间 $\mathcal{F}$ 中 $\hat{f}_n \xrightarrow{P} f_0$. 因此在空间 Π 中 $(\mathbb{G}_n, \hat{f}_n) \rightsquigarrow (\mathbb{G}_P, f_0)$. 由引理 A.20, $\mathbb{G}_P$ 的几乎所有样本轨道在 $\mathcal{F}$ 上连续, 因而函数 g 在几乎每个点 $(\mathbb{G}_P, f_0)$ 处连续. 由连续映照定理, $\mathbb{G}_n(\hat{f}_n - f_0) = g(\mathbb{G}_n, \hat{f}_n) \rightsquigarrow g(\mathbb{G}_P, f_0) = 0$. 这就证明了引理, 因为在极限退化为 0 时弱收敛即为依概率收敛. ■

A.4.4 $\mathcal{F}$ 改变为 $\mathcal{F}_n$ 的情形

对经验过程弱收敛的研究, 往往需要突破函数类 $\mathcal{F}$ 不随 n 改变的约束. 例如 M-估计的渐近研究就提出了这类要求. 另一方面, 即使做了这样的推广, 有关定理仍然可以借助于极大不等式推出.

设 $\mathcal{F}_n$ 由一类可测函数 $f_{n,t} : \mathcal{X} \mapsto \mathbf{R}$ 构成, 其中 t 属于一个共同的指标集 T. 则我们可以考虑经验过程 $t \mapsto \mathbb{G}_n f_{n,t}$ 作为 $l^\infty(T)$ 的成员的弱收敛, 此处我们假定样本轨道是有界的. 由定理 A.19, 弱收敛等价于边缘收敛加上渐近胎紧. 到 Gauss 过程的边缘收敛可由 Lindeberg 条件得出, 渐近胎紧则可由 $\mathcal{F}_n$ 的熵的有关条件得出.

假定在度量 ρ 之下 T 为一全有界空间, 且相对于 L_2-距离, 有

$$\sup_{\rho(s,t)<\delta_n} P(f_{n,s} - f_{n,t})^2 \to 0, \quad \text{对每个 } \delta_n \downarrow 0. \tag{A.5}$$

进一步, 假定函数类 $\mathcal{F}_n$ 的包络函数 F_n 满足 Lindeberg 条件

$$PF_n^2 = O(1),$$

$$PF_n^2 I\{F_n > \epsilon\sqrt{n}\} \to 0, \quad \text{对每个 } \epsilon > 0.$$

关于经验过程序列 $\{\mathbb{G}_n f_{n,t} : t \in T\}$ 的弱收敛, 我们有下面的定理.

定理 A.28 设 $\mathcal{F}_n = \{f_{n,t} : t \in T\}$ 是一类可测函数, (T, ρ) 为一全有界距离空间, 满足 (A.5), 且 Lindeberg 条件成立. 假定对每个 $\delta_n \downarrow 0$, 有 $J_{[]}\big(\delta_n, \mathcal{F}_n, L_2(P)\big) \to 0$. 则序列 $\{\mathbb{G}_n f_{n,t} : t \in T\}$ 弱收敛于一个胎紧的 Gauss 过程, 只要协方差函数序列 $Pf_{n,s}f_{n,t} - Pf_{n,s}Pf_{n,t}$ 在 $T \times T$ 上点点收敛.

证明 与定理 A.24 类似, 只要证明序列 $\{\mathbb{G}_n f_{n,t} : t \in T\}$ 的渐近胎紧性, 即证明定理 A.19 中的 (ii). 由 (A.5) 及距离空间 (T, ρ) 的全有界性, 我们可将 T 分割为有限多个子集 $T_1, \cdots, T_k$, 使得对充分大的 n,

$$\sup_i \sup_{s,t \in T_i} P(f_{n,s} - f_{n,t})^2 < \delta^2.$$

(这是唯一用到 T 的全有界性的地方, 当然我们也可直接假定存在这样的分割.)

由极大不等式可得

$$\begin{aligned} &\mathrm{E}^*\Big[\sup_i \sup_{s,t \in T_i} |\mathbb{G}_n(f_{n,s} - f_{n,t})|\Big] \\ \lesssim\ & J_{[]}\big(\delta, \mathcal{F}_n, L_2(P)\big) + PF_n^2 I\{F_n > a_n(\delta)\sqrt{n}\}/a_n(\delta). \end{aligned}$$

此处, 由引理 A.25 得出的不等式, 以及 $a_n(\delta)$ 的计算, 都已考虑到有关的函数类为 $\mathcal{G}_n := \mathcal{F}_n - \mathcal{F}_n$ 而包络为 $2F_n$, 但与函数类 $\mathcal{F}_n$ 相比, 无论是包络、熵积分还是 $a_n(\delta)$, 都仅仅相差一个常数因子. 由假定, 对每个 $\delta_n \downarrow 0$, 有 $J_{[]}\big(\delta_n, \mathcal{F}_n, L_2(P)\big) \to 0$, 故对每个给定的 $\delta > 0$, $J_{[]}\big(\delta, \mathcal{F}_n, L_2(P)\big) = O(1)$ 因而 $a_n(\delta)$ 有个大于 0 的下界. 由 Lindeberg 条件, 对每个给定的 $\delta > 0$, 上式右方的第二项趋于 0. 由假定, 可选取 δ 充分小, 使得当 n 充分大时第一项也任意小. 由 Markov 不等式, 即知渐近胎紧性成立因而定理得证. ■

在 M-估计的研究中, 在处理有关随机函数列的收敛问题时, 有时需要对给定函数类 "重新刻度化". 今设 $x \mapsto m_\theta(x)$, $\theta \in \Theta$ 为一函数类, 其中参数集 Θ 为欧氏的. 我们需要给出合适的条件, 使得对任意给定的序列 $r_n \to \infty$, 任一给定的常数 $c > 0$, 和任意的随机序列 $\tilde{h}_n = O_P^*(1)$,

$$\sup_{||h|| \leqslant c} \Big|\mathbb{G}_n\Big(r_n[m_{\theta_0 + h/r_n} - m_{\theta_0}] - h^\tau \dot{m}_{\theta_0}\Big)\Big| \xrightarrow{P} 0. \tag{A.6}$$

$$\mathbb{G}_n\Big(r_n[m_{\theta_0 + \tilde{h}_n/r_n} - m_{\theta_0}] - \tilde{h}_n^\tau \dot{m}_{\theta_0}\Big) \xrightarrow{P} 0. \tag{A.7}$$

定理 A.29 设下面涉及的每个 θ 都在 Θ 的内部, $x \mapsto m_\theta(x)$ 为一可测函数. 映照 $\theta \mapsto m_\theta(x)$ 对几乎每个 x (或依概率) 在 θ_0 可微, 偏导数记为

$$\dot{m}_{\theta_0}(x) := \frac{\partial m_\theta(x)}{\partial \theta}\Big|_{\theta = \theta_0}.$$

假定存在一个 P 平方可积函数 K, 和 θ_0 的一个邻域, 对其中的每对 θ_1 和 θ_2 都有

$$|m_{\theta_1}(x) - m_{\theta_2}(x)| \leqslant K(x)||\theta_1 - \theta_2||.$$

则对任一给定的常数 $c > 0$, 和依概率有界的任一随机序列 $\tilde{h}_n$, (A.6) 和 (A.7) 成立.

证明 由映照 $\theta \mapsto m_\theta(x)$ 的依概率可微性和 Lipschitz 条件, 应用控制收敛定理, 知随机变量序列

$$g_n(h,\omega) := \mathbb{G}_n\Big(r_n[m_{\theta_0+h/r_n} - m_{\theta_0}] - h^\tau \dot{m}_{\theta_0}\Big)$$

均值为 0, 方差收敛到 0, 因而该序列依概率收敛于 0. 且 $\dot{m}_{\theta_0} \in L_2(P)$. 不妨设 $||\dot{m}_{\theta_0}||_{P,2} \leqslant ||K||_{P,2}$. 换句话说, 若将 $g_n(h,\omega)$ 视为以 h 为指标参数的过程序列, 其边缘依分布收敛于 0.

设 $\Theta \subset \mathbf{R}^k$. 以 $S_1 = \{||h|| \leqslant 1\}$ 记 $\mathbf{R}^k$ 内的单位球. 今若知道过程序列在空间 $l^\infty(S_1)$ 上弱收敛, 则由边缘收敛知过程的弱极限必为一个恒等于 0 的过程, 因 $z \mapsto \sup_{h\in S_1}|Z(h)|$ 为空间 $l^\infty(S_1)$ 上的连续函数, 由连续映照定理, 即得 $\sup_{h\in S_1}|g_n(h,\omega)| \rightsquigarrow 0$, 或 $\sup_{h\in S_1}|g_n(h,\omega)| \xrightarrow{P} 0$, 此即 (A.6). 由此及序列 $\tilde{h}_n$ 依概率有界, 即得 (A.7). 因此, 问题归结为证明过程的弱收敛.

上述过程可视为以函数类 $\mathcal{F}_n := r_n\mathcal{M}_{1/r_n}$ 为指标集的经验过程, 其中 $\mathcal{M}_\delta := \{f_\theta : ||\theta-\theta_0|| \leqslant \delta\}$, $f_\theta(x) = m_\theta(x) - m_{\theta_0}(x) - (\theta-\theta_0)^\tau \dot{m}_{\theta_0}(x)$. 由 Lipschitz 条件, 函数 $M_\delta = 2\delta K$ 为 $\mathcal{M}_\delta$ 的包络函数.

由例 A.4, 函数类 $\mathcal{M}_\delta$ 的括号数满足

$$N_{[]}(\epsilon\delta||2K||_{P,2}, \mathcal{M}_\delta, L_2(P)) \leqslant C\epsilon^{-k}, \quad 0<\epsilon<\delta.$$

常数 C 与 ϵ 及 δ 无关. 由此可算得以 $M_\delta/\delta = 2K$ 为包络的函数类 $\mathcal{M}_\delta/\delta$ 的括号数, 因而有

$$J_{[]}(\delta_n, \mathcal{M}_\delta/\delta, L_2(P)) \lesssim \int_0^{\delta_n} \big[k\mathrm{Log}(1/\epsilon) + \mathrm{Log}C\big]^{1/2} d\epsilon,$$

故当 $\delta_n \downarrow 0$ 时上式左方关于 δ 一致地收敛于 0; 再加上包络函数 $M_\delta/\delta = 2K$ 满足 Lindeberg 条件, 由定理 A.28 即知定理成立. ■

参考文献

[1] Andersen, E. B. (1970). Sufficiency and exponential families for discrete sample spaces. *J. Amer. Statit. Assoc.*, **65**, 1248–1255.

[2] Andersen, T. W. and Darling, D. A. (1952). Asymptotic theory of certain "goodness of fit" criteria based on stochastic processes. *Ann. Math. Statist.*, **23**, 193–212.

[3] Bahadur, R. R. (1957). On unbiased estimates of uniformly minimum variance. *Sankhyā*, **18**, 211–224.

[4] Bahadur, R. R. (1966). A note on quantiles in large sample. *Ann. Math. Statist.*, **37**, 577–580.

[5] Bai, Z. D., Chen, X. R., Wu, Y. H. and Zhao, L. C. (1987). Asymptotic normality of minimum L_1-norm estimates in linear models. Technical Report, 87–35, Center for Multivariate Analysis, University of Pittsburgh.

[6] Bai, Z. D., Rao, C. R. and Wu, Y, (1992). M-estimation of multivariate linear regression parameters under a convex discrepancy function. *Statistica Sinica*, **2**, 237–254.

[7] Barndorff-Nielsen, O. and Pedersen, K. (1968). Sufficient data reduction and exponential families. *Math. Scand.*, **22**, 197–202.

[8] Bartlett, M. S. (1937). Properties of sufficiency and statistical tests. *Proc. Roy. Sec. London, Ser. A*, **160**, 268–282.

[9] Basu, D. (1955). On statistics independent of a complete sufficient statistic. *Sankhyā*, **15**, 377–380.

[10] Basu, D. (1958). On statistics independent of sufficient statistics. *Sankhyā*, **20**, 223–226.

[11] Bayes, T. (1763). An essay tiward solving a problem in the doctrine of chances. *Phil. Trans. Roy. Soc.*, **153**, 370–418. Reprinted in *Biometrika*, 1958, **45**, 293–315.

[12] Bennett, G. (1962). Probability inequalities for the sums of independent random variables. *J. Amer. Statist. Assoc.*, **57**, 33–45.

[13] Berger, J. O. (1985). *Statistical Decision Theory and Bayesian Analysis*, 2nd ed., Springe-Verlag, New York.

[14] Billingsley, P. (1986). *Convwrgence of probability Measures*, 1st ed., Wiley, New York.

[15] Billingsley, P. (1999). *Convwrgence of probability Measures*, 2nd ed., Wiley, New York.

[16] Blackwell, D. (1947). Conditional expectation and unbiased sequential estimation. *Ann. Math. Statist.*, **18**, 105–110.

[17] Box, G. E. and Tiao, G. C. (1973). *Bayesian Inference in Statistical Analysis.* Addison-Wesley. Reading.

[18] Chen, X. R. (1979). Consistency of least squares estimates in linear models. *Scientia Sinica*, Special Issue (II), 162–176.

[19] Chen, X. R., Bai, Z. D., Zhao, L. C. and Wu, Y. H. (1990). Asymptotic normality of minimum L_1-norm estimates in linear models. *Science in China*, Ser. A, Chinese Edition: **20**, 162–177; English Edition: **33**, 1311–1328.

[20] Chen, X. R., Bai, Z. D., Zhao, L. C. and Wu, Y. H. (1992). Consistency of minimum L_1-norm estimates in linear models. In *The development of Statistics: Recent contributions from China* (eds. Chen, X. R., Fang, K. T., Yang, C. C.), Pitman Research Notes in Mathematics Series 258, Longman Scientific & Technical, 249–260.

[21] Chen, X. R. and Wu, Y. (1994). Nonexistence of consistent estimates in a density estimation problem. *Statist. Probab. letters*, **21**, 141–145.

[22] Chernoff, H. (1954). On the distribution of the likelihood test statistic. *Ann. Math. Statist.*, **25**, 579–586.

[23] Chung, K. L. (1974). *A Course in Probability Theory*, 2nd ed., Academic Press, New York.

[24] Coles, S. (2007). *An Introduction to Statistical Modeling of Extreme Values* (极值统计建模导论). Springer, 世界图书出版公司 (2008).

[25] Cramér, H. (1928). On the composition of elementrary errors. *Skand. Aktuarietidskr.*, **11**, 13–74, 141–186.

[26] Cramér, H. (1946a). *Mathematical Methods of Statistics.* Princeton University Press, Princeton.

[27] Cramér, H. (1946b). A contribution to the theory of statistical estimation. *Skand. Akt. Tidskr.*, **29**, 85–94.

[28] Doob, J. L. (1953). *Stochastic Processes.* Wiley, New York.

[29] Drygas, H. (1976). Weak and strong consistency of the least squares estimators in regression model. *Z. Wahrsch. Verw. Gabiete*, **34**, 119–127.

[30] Durbin, J. (1973a). *Distribution Theory for Tests Based on the Sample Distribution Function.* SIAM, Philadelphia, PA.

[31] Durbin, J. (1973b). Weak convergence of the sample distribution function when parameters are estimated. *Ann. Statist.*, **1**, 279–290.

[32] Edgeworth, F. Y. (1908, 1909). On the probable errors of frequency constants. *J. Roy. Statist. Soc. Ser. B*, **71**, 381–397, 499–512, 651–678; **72**, 81–90.

[33] Fahmeir, L. and Tutz, G. (1994). *Multivariate Statistical Modelling Based on Generalized Linear Models*. Springer, 世界图书出版公司 (1998).

[34] Fang, K. T., Yang, Z. H. and Kotz, S. (2001). Generation of multivariate distributions by vertical density representation. *Statistics*, **35**, 281–293.

[35] Farrell, R. H. (1966). Weak limits of sequence of Bayes procedures in estimation theory. *Proc. Fifth Berkeley Symp. Math. Statist. Probab.*, **1**, 83–111. University of California Press, Berkeley.

[36] Feller, W. (1957). *An Introduction to Probability Theory and Its Applications*. Vol. 1, 2nd ed., Wiley, New York. [中译本见后面]

[37] Ferguson, T. S. (1996). *A Course in Large Sample Theory*. Chapman & Hall, London.

[38] de Finetti, B. (1937). La prévision: Ses lois logiques, ses source subjectives. *Ann. de l'Inst. Henri Poincaré*, **7**, 1–68. (Translated in *Studies in Subjective Probability*. Kyburg and Smokler, Eds. Wiley, New York.

[39] de Finetti, B. (1970). *Teoria Delle Probabilità*. English Translation(1974) *Theory of Probability*. Wiley, New York.

[40] Fisher, R. A. (1920). A mathematical examination of the methods of determining the accuracy of an observation by the mean error, and by the mean square error. *Monthly Notices R. Astron. Soc.*, **80**, 758–770.

[41] Fisher, R. A. (1922). On the mathematical fundations of the theoretical statistics. *Philos. Trans. Roy. Soc. London, Ser. A*, **222**, 309–368.

[42] Fisher, R. A. (1924). The conditions under which chi square measures the discrepancy between observation and hypothesis. *J. Roy. Statist. Soc.*, **87**, 442–450.

[43] Fisher, R. A. (1925). Theory of statistical estimation. *Proc. Camb. Phil. Soc.*, **22**, 700–725.

[44] Fisher, R. A. (1930). Inverse probability. *Proc. Cambridge Philos. Soc.*, **26**, 528–535.

[45] Fisher, R. A. (1935). The fiducial argument in statistical inference. *Ann. Eugen.*, **6**, 391–398.

[46] Fisher, R. A. (1973). *Statistical Methods and Scientific Inference*, 3rd., Hafner, New York.

[47] Fisher, R. A. and Yates, F. (1963). *Statistical Tables for Biological, Agricultural Medical Research*, 6th Ed., Oliver and Boyd, Edinburgh and London.

[48] Gauss, C. F. (1821). Theoria combinationis observationum erroribus minimis obnoxiae.[An English translation can be found in Gauss's work (1803—1826) on the Theory of Least Squares. Tech. Rep. No. 5. Princeton University. Princeton.]

[49] Girshick, M. A. and Savage, L. J. (1951). Bayes and minimax estimates for quadratic loss functions. University of California Press, 53–73.

[50] Gnedenko, B. (1943). Sur la distribution limit du terme maximum d'une série aléatoire. *Ann. Math.*, **44**, 423–453.

[51] Gumbel, E. J. (1958). *Statistics of Extremes.* Columbia Univ. Press, New York.

[52] de Haan, L. (1970). *On Regular Variation and its Application to the Weak Convergence of Sample Extremes.* Math. Centre Tracts 32, Amsterdam.

[53] de Haan, L. (1976). Sample Extremes: An elementary introduction. *Statist. Neerlandica*, **30**, 161–172.

[54] Hahn, G. J. and Hendrickson, R. W. (1971). A table of percentage points of the distribution of the largest absolute value of k Student t variates and its applications. *Biometrika*, **58**, 323–332.

[55] Hájek, J. (1962). Asymptotically most powerful rank order tests. *Ann. Math. statst.*, **33**, 1124–1147.

[56] Hájek, J. (1968). Asymptotic normality of simple linear rank statistics under alternatives. *Ann. Math. statst.*, **39**, 325–346.

[57] Hájek, J. and Šidák, Z. (1967). *Theory of rank tests.*, Academic Press, New York.

[58] Hájek, J. (1972). Local asymptotic minimax and admissibility in estimation. In *Proc. 6th Berkeley Symp. on Math. Statist. and Probab.*, **1**, 175–194. Univ. of California Press, Berkeley.

[59] Hall, P. and Heyde, C. C. (1980). *Martingale Limit Theory and Its Application*, Academic Press, New York.

[60] Halmos, P. R. (1950). *Measure Theory*, Van Nostrand, New York. (中译本) 测度论, 王建华译, 科学出版社, 北京, 1965.

[61] Halmos, P. R. and Savage, L. J. (1949). Application of the Radon-Nikodym theorem to the theory of sufficient statistics. *Ann. Math. Statist.*, **20**, 225–241.

[62] Harter, H. L. (1960). Tables of range and Studentized range. *Ann. Math. Statist.*, **31**, 1122–1147.

[63] Hodges, J. L., Jr., and Lehmann, E. L. (1950). Some problems in minimax point estimation. *Ann. Math. Statist.*, **21**, 182–197.

[64] Hoeffding, W. (1963). Probability inequalities for sums of bounded random variables. *J. Amer. Statist. Assoc.*, **58**, 13–30.

[65] Huber, P. J. (1964). Robust estimation of a location parameter. *Ann. Math. Statist.*, **35**, 73–101.

[66] Huber, P. J. (1973). Robust regression. *Ann. Statist.*, **1**, 799–821.

[67] Huber, P. J. (1981). *Robust Regression.* Wiley, New York.

[68] James, W. and Stein, C. (1961). Estimation with quadratic loss. *Proc. Fourth Berkeley Symp. Math. Statist. Probab.*, **1**, 311–319. University of California Press, Berkeley.

[69] Jeffreys, H. (1939, 1948, 1961). *The Theory of Probability.* Oxford University Press, Oxford.

[70] Johnson, N. L. and Kotz, S. (1970a). *Continuous Univariate Distributions-I*. Houghton Mifflin, Boston.

[71] Johnson, N. L. and Kotz, S. (1970b). *Continuous Univariate Distributions-II*. Houghton Mifflin, Boston.

[72] Jurečková, J. and Sen, P. K. (1996). *Robust Statistical Procedures: Asymptotics and Interrelations*. Wiley, New York.

[73] Karlin, S. (1958). Admissibility for estimation with quadratic loss. *Ann. Math. Statist.*, **29**, 406–436.

[74] Kendall, M. G. and Stuart, A. (1977). *Advanced Theory of Statistics*. Vol. 1, Griffin, London.

[75] Kendall, M. G. and Stuart, A. (1979). *Advanced Theory of Statistics*. Vol. 2, Griffin, London.

[76] Kendall, M. G. and Stuart, A. (1976). *Advanced Theory of Statistics*. Vol. 3, Griffin, London.

[77] Kiefer, J. (1957). Invariance, minimax sequential estimation, and continuous time processes. *Ann. Math. Statist.*, **28**, 573–601.

[78] Kiefer, J. (1967). On Bahadur's representation of sample quantiles. *Ann. Math. Statist.*, **38**, 1323–1342.

[79] Kiefer, J. (1970a). Deviations between the sample quantiles process and the sample df. *Proc. Conference on Nonparametric Techniques in Statistical Inference*, Bloomington, ed. by M. L. Puri, Cambridge University Press, 299–319.

[80] Kiefer, J. (1970b). Old and new methods for studying order statistics and sample quantiles. **ibid**., 349–357.

[81] Kolmogorov, A. (1933). Sulla adeterminazione empirica di una legge di distribuzione. *Giorn. Inst. Ital. Attuari*, **4**, 83–91.

[82] Knight, K. (1999). *Mathematical Statistics*. Chapman & Hall/CRC, Boca Raton.

[83] Lai, T. L., Robbins, H. and Wei, C. Z. (1979). Strong consistency of least squares estimates in multiple regression II. *J. Multivariate Anal.*, **9**, 343–362.

[84] Laplace, P. S. (1774). Mémoire sur la probabilité des causes par évènements. *Mém. Acad. Sci. Sav. Etranger*, **6**, 621–656.

[85] Laplace, P. S. (1812). *Théorie Analytique des Probabilités*. Courcier, Paris.

[86] Laplace, P. S. (1820). *Theorie Analytique des Probabilités*, 3rd ed. Courcier, Paris.

[87] Le Cam, L. (1953). On some asymptotic properties of maximum likelihood estimates and related Bayes estimates. *Univ. of Calif. Publ. in Statist.*, **1**, 277–330.

[88] Le Cam, L. (1960). Locally asymptotically normal families of distributions. *Univ. of Calif. Publ. in Statist.*, **3**, 37–98.

[89] Le Cam, L. (1969). *Theorie Asymptotique de la Decision Statistique*. Presses de l'Université de Motreal.

[90] Le Cam, L. (1970). On the assumption used to prove asymptotic normality of maximum likelihood estimates. *Ann. Math. Statist.*, **41**, 802–828.

[91] Le Cam, L. (1979). On a theorem of J. Hájek. In *Contribution to Statistics: J. Hájek Memorial Volume* (Jureckova, ed.), Academia, Prague.

[Rigorous and very general treatment of the large-sample theory of maximum likelihood estimates, with a survey of the large previous literature on the subject.]

[92] Le Cam, L. (1972). Limits of experiments. *Proc. 6th Berkeley Symp. on Math. Statist. and Probab.* **1**, 245–261. Univ. of California Press, Berkeley.

[93] Le Cam, L. (1986). *Asymptotic Methods in Statistical Decision Theory.* Springer-Verlag, New York.

[94] Le Cam, L. and Yang, G. (1990). *Asymptotics in Statistics, Some Basic Concepts*, 1st ed. Springer-Verlag, New York. [中文译本见勒康, 罗昭容 (1993).] 2nd ed. (2000).

[95] Lehmann, E. L. (1983). *Theory of Point Estimation.* Wiley, New York.

[96] Lehmann, E. L. (1986). *Testing Statistical Hypotheses*, 2nd ed. Wiley, New York.

[97] Lehmann, E. L. and Casella, G. (1998). *Theory of Point Estimation*, 2nd. Springer, New York. (中译本) 点估计理论, 2 版, 郑忠国, 蒋建成, 童行伟译, 中国统计出版社, 2005.

[98] Lehmann, E. L. and Romana, J. P. (2005). *Testing Statistical Hypotheses.* 3rd ed. Springer.

[99] Lehmann, E. L. and Scheffé, H. (1950, 1955, 1956). Completeness, similar regions, and unbiased estimation. *Sankhyā*, **10**, 305–340; **15**, 219–236; correction **17**,250.

[100] Lindley, D. V. (1965). *Introduction to Probalibility and Statistics from a Bayesian Viewpoint. Part 2, Inference.* Cambridge University Press, Cambridge.

[101] Madansky, A. (1962). More on length of confidence intervals. *J. Amer. Statist. Assoc.*, **57**, 586–589.

[102] Manski, C. (1975). Maximum score estimation of the stochastic utility model of choice. *J. Econometrics*, **3**, 205–228.

[103] Manski, C. (1985) Semiparametric analysis discrete response, asymptotic properties of the maximum score estimator. *J. Econometrics*, **27**, 313–333.

[104] Munkres, J. R. (1991). *Analysis on Manifolds.* Westview Press.

[105] Neyman, J. (1937). Outline of a theory of statistical estimation based on the classical theory of probability. *Phil. Trans. Roy. Soc., Ser. A*, **236**, 333–380.

[106] Neyman, J. and Pearson, E. S. (1928). On the use and interpretation of certain test criteria. *Biometrika*, **20A**, 175–240, 263–294.

[107] Neyman, J. and Pearson, E. S. (1933). On the problem of the most efficient tests of statistical hypotheses. *Phil. Trans. Roy. Soc. Ser. A*, **231**, 289–337.

[108] Neyman, J. and Pearson, E. S. (1936a). Contributions to the theory of testing statistical hypotheses. I. Unbiased critical region of type A and type A_1. *Statist.Rev. Mem.*, **1**, 1–37.

[109] Neyman, J. and Pearson, E. S. (1936b). Sufficient statistics and uniformly most powerful tests of statistical hypotheses. *Statist.Rev. Mem.*, **1**, 113–137.

[110] Owen, A. B. (1988). Empirical likelihood ratio confidence intervals for a single functional. *Biometrika*, **75**, 237–249.

[111] Owen, A. B. (1990). Empirical likelihood confidence regions. *Ann. Statist.*, **18**, 90–120.

[112] Owen, A. B. (2001). *Empirical Likelihood*. Chapman & Hall/CRC, Boca Raton.

[113] Pearson, K. (1900). On the criterion that a given system of deviations from the probable in the case of a correlated system of variables is such that it can be reasonably supposed to have arisen from random sampling. *Philosophical Magazine, Ser.5*, **50**, 157–175. (Reprinted in: *Karl Pearson's Early Statistical Papers*, Cambridge University Press, 1956).

[114] Peisakoff, M. (1950). Transformation parameters. Unpublished Ph. D. thesis. Princeton University, Princeton.

[115] Pitman, E. J. G. (1939). Tha estimation of the location and scal parameters of a continuous population of any given form. *Biometrika*, **30**, 391–421.

[116] Pollard, D. (1990). *Empirical Procecess: Theory and Application*. NSF-CBMS Regional Conference Series in Probability and Statistics, Vol. 2. Institute of Mathematical Statistics, Hayward.

[117] Pollard, D. (1991). Asympototics for least absolute deviation regression estimators. *Econometric Theory*, **7**, 186–199.

[118] Powell, J. L. (1984). Least absolute deviations estimates for the censored regression model. *J. Econometrics*, **25**, 303–325.

[119] Powell, J. L., Stock, J. H. and Stocker, T. M. (1989). Semiparametric estimation of index coefficients. *Econometrica*, **57**, 1403–1430.

[120] Raiffa, H. and Schlaifer, R. (1961). *Applied Statistical Decision Theory*. Harvard University Press, Cambrige.

[121] Rao, C. R. (1945). Information and the accuracy attainable in the estimation of statistical parameters. *Bull. Calc. Math. Soc.*, **37**, 81–91.

[122] Rao, C. R. (1992). R. A. Fisher: The founder of mogern statistics. *Statistical Science*, **7**, 34–48.

[123] Rao, C. R. and Toutenburg, H. (1995). *Linear Models: Least Squares and Alternatives*. Springer, New York.

[124] Rao, C. R. and Zhao, L. C. (1992). Approximation to the distribution of M-estimates in linear models by randomly weighted bootstrap. *Sankhyā*, Ser. A, **54**, 323–331.

[125] Rao, C. R. and Zhao, L. C. (1993). Asymptotic normality of LAD estimator in censored regression models. *Mathematical Methods of Statistics*, **2**, 228–239.

[126] Rao, C. R. and Zhao, L. C. (1995). Recent contributions to censored regression models. *Metrika*, **42**, 203–213.

[127] Reiss, R. -D. (1989). *Approximate Distributions of Order Statistics with Applications to Nonparametric Statistics*. Springer, New York.

[128] Resnick, S. I. (1987). *Extreme Values, Regular Variation, and Point Processes*. Springer, New York.

[129] Rubin, D. B. (1984). Bayesianly justtifiable and revelant frequency calculations for the applied statistician. *Ann. Statist.*, **12**, 1151–1172.

[130] Sacks, J. (1963). Generalized Bayes solution in estimation problems. *Ann. Math. Statist.*, **10**, 643–645.

[131] Savage, L. J. (1954). *The Foundations of Statistics*. Wiley, New York.

[132] Serfling R. J. (1980). *Appriximation Theorems of Mathematical Statistics*. Wiley, New York

[133] Shao, Jun (2003). *Mathematical Statistics*, 2nd ed., Springer, 世界图书出版公司 (2009).

[134] Shao, Jun (2005). *Mathematical Statistics: Exercises and Solutions*, Springer, 世界图书出版公司 (2010).

[135] Smirnov, N. V. (1939). An estimate of divergence between empirical curves of a distribution in two independent samples. *Bull. MGU*, **2**, 3–14 (in Russian).

[136] Smirnov, N. V. (1944). Approximate laws of distribution of random variables from empirical data. *Usp. Mat. Nauk*, **10**, 179–206 (in Russian).

[137] Stein, C. (1955). A necessary and sufficient condition for admissibility. *Ann. Math. Statist.*, **26**, 518–522.

[138] Stein, C. (1956). Inadmissibility of the usual estimator for the mean of a multivariate distribution. *Proc. Third Berkeley Symp. Math. Statist. Prob.*, **1**, University of California Press, 197–206.

[139] Stein, C. (1959). The admissibility of Pitman's estimator for a single location parameter. *Ann. Math. Statist.*, **30**, 970–979.

[140] Stein, C. (1961). Inadmissibility of the usual estimator for the mean of a multivariate distribution. *Proc. Third Berkeley Symp. Math. Statist. Prob.*, **1**, 197–206. University of California Press, Berkeley.

[141] Stein, C. (1964). Inadmissibility of the usual estimate for the variance of a normal distribution with unknown mean. *Ann. Inst. Statist. Math.*, **16**, 155–160.

[142] Stigler, S. M. (1969). Linear functions of order statistics. *Ann. Math. Statist.*, **40**, 770–788.

[143] Stigler, S. M. (1973). The asymptotic distribution of the trimmed mean. *Ann. Statist.*, **1**, 472–477.

[144] Stigler, S. M. (1974). Linear functions of order statistics with smooth weight functions. *Ann. Statist.*, **2**, 676–693; (1979) Correction note. *Ann. Statist.*, **7**, 466.

[145] Tate, R. F. (1959). Unbiased estimation: function of location and scale parameters. *Ann. Math. Statist.*, **30**, 341–366.

[146] Troutt, M. D. (1991). A theorem on the density of ordinate and an alternative derivation of the Box-Muller method. *Statistics*, **22**, 436–466.

[147] van der Vaart, A. W. and Wellner, J. A. (1996). *Weak Convergence and Empirical Processes*. Springer, New York.

[148] van der Vaart, A. W. (1998). *Asymptotic Statistics*. Cambridge University Press, Cambridge.

[149] Wald, A. (1949). Note on the consistency of the maximum likelihood estimate. *Ann. Math. Statist.*, **20**, 595–601.

[150] Wald, A. (1950). *Statistical Decision Functions*. Wiley, New York.

[151] Weissman, I. (1975). Extremal processes generated by independent nonidentically distributed random variables. *Ann. Probability*, **3**, 172–178.

[152] Wilks, S. S. (1938). The large-sample distribution of the likelihood ratio for testing composite hypotheses. *Ann. Math. Statist.*, **9**, 60–62.

[153] Wu, Y. H. and Zhao, L. C. (1999). A large sample study of randomly weighted bootstrap in linear models. *Science in China*, Ser. A, **42**, 1066–1074.

[154] Xie, M. G. and Singh, K. (2013). Confidence distribution, the frequentist distribution estimator of a parameter: a review. *International Statistical Review*, **81**, 3–39.

[155] Zacks, S. (1971). *Theory of Statistical Inference*. Wiley, New York.

[156] Zhao, L. C. (2000). Some contributions to M-estimation in linear models. *J. Statist. Plann. Inference.*, **88**, 189–203.

[157] Zhao, L. C. (2002). Strong consistency of M-estimates in linear models. *Science in China*, Ser. A, **45**, 1420–1427.

[158] Zhao, L. C., Rao, C. R. and Chen, X. R. (1993). A note on the consistency of M-estimates in linear models. In *Stochastic Process, A Festschrift in Honour of Gopinath Kallianpur* (eds. Cambanis, S., Ghosh, J. K., Karandikar, R. L., Sen, P. K.), Springer-Verlag, 359–367.

[159] Zhao, L. C. (2004). Linear hypothesis testing in censored regression models. *Statistica Sinica*, **14**, 333–347.

[160] 成平, 陈希孺, 陈桂景, 吴传义 (1985). 参数估计. 上海科学技术出版社, 上海.

[161] 陈桂景, 陈希孺 (1982). 同变性限制下的充分性原则. 应用数学学报, 第 5 卷, 第 1 期, 85–93.

[162] 陈家鼎, 孙山泽, 李东风, 刘力平 (2006). 数理统计学讲义. 2 版. 高等教育出版社, 北京.

[163] 陈希孺 (1980). 关于 U 统计量和 von-Mises 统计量的极限性质. 中国科学, 第 10 卷, 第 6 期, 522–532.

[164] 陈希孺 (1981, 1997). 数理统计引论. 科学出版社, 北京.

[165] 陈希孺 (1999, 2009). 高等数理统计学. 中国科学技术大学出版社, 合肥.

[166] 陈希孺, 倪国熙 (1988). 数理统计学教程. 上海科学技术出版社, 上海.

[167] 陈希孺, 赵林城 (1996). 线性模型中的 M 方法. 上海科学技术出版社, 上海.

[168] 费勒, W., (1964). 概率论及其应用. 上册. 胡迪鹤, 林向清译. 科学出版社, 北京.

[169] 贾柯勃逊, N. (1960). 抽象代数学. 卷 1-基本概念. 黄缘芳译. 科学出版社, 北京.

[170] 勒康, 罗昭容 (1993). 统计渐近论基础. 科学出版社, 北京.

[171] 卢昆亮, 赵林城 (1982). 双边截断参数的最小方差无偏估计. 应用数学学报, 第 5 卷, 第 2 期, 177–185.

[172] 茆诗松, 王静龙, 濮晓龙 (1998). 高等数理统计. 高等教育出版社, 北京.

[173] 倪国熙 (1984). 常用的矩阵理论和方法. 上海科学技术出版社, 上海.

[174] 史宁中 (2008). 统计检验的理论与方法. 科学出版社, 北京.

[175] 韦博成 (2006). 参数统计教程. 高等教育出版社, 北京.

[176] 韦来生, 张伟平 (2013). 贝叶斯分析. 中国科学技术大学出版社, 合肥.

[177] 严加安 (1981). 鞅与随机积分引论. 上海科学技术出版社, 上海.

[178] 严士健, 刘秀芳 (1994). 测度与概率. 北京师范大学出版社, 北京.

[179] 杨振海, 程维虎, 张军舰 (2011). 拟合优度检验. 科学出版社, 北京.

[180] 杨振海 (2014). 随机估计及 VDR 检验. 高等教育出版社, 北京.

[181] 张里千 (1955). 论斯米尔诺夫统计量的真确分布及其渐近展开. 数学进展, **1**, 775–790.

[182] 张里千 (1956). 论柯尔莫哥洛夫统计量的真确分布及其渐近展开. 数学学报, **6**, 55–81.

[183] 郑忠国, 童行伟, 赵慧 (2012). 高等统计学. 北京大学出版社, 北京.

索引

F-Brown 桥, 254
F-分布, 25
F-检验, 124
L-统计量, 228
M-估计, 179, 232, 259
M-估计的分布的随机加权逼近, 234
$O_p(1)$, $o_p(1)$, 249
P-Donsker 类, 255
P-Glivenko-Cantelli 类, 255
$X \sim P_\theta$, 12
χ^2-分布, 25
χ^2-检验, 123
χ^2-拟合优度检验, 205
χ^2-统计量, 205
t-分布, 25
t-检验, 123
t-区间估计, 138
0 的无偏估计, 56

a.e., 4, 75
a.s. 收敛, 247, 252

Bartlett 修正, 203
Bayes 估计, 68
Bayes 可信区间, 149
Bayes 解, 69
Bayes 可信区域, 149
Bayes 区间估计, 149
Bayes 统计决策问题, 68
Bayes 统计推断, 83
Bayes 准则, 22
Beinstein 不等式, 82
Brown 桥, 254

Cramér-Wold device, 249
Cramér-Rao 不等式, C-R 不等式, 56
Cramér-von Mises 检验, 212

D'Agostino 的 D 检验, 211
Delta 方法, 165
Donsker 定理, 254

Fisher 信息函数, 59
Fisher 信息矩阵, 61
Fubini 定理, 4

Gauss 过程, 254, 255, 259
Glivenko-Cantelli 定理, 254

Haar 测度, 89
Helley 选择定理, 249

Jensen 不等式, 40

Kolmogorov 检验, 211
Kolmogorov 统计量, 24
Kronecker 引理, 168
Kullback-Leibler 信息量, 40

Le Cam 第三引理, 198
Le Cam 第一引理, 196
Lehmann-Scheffé 定理, 50
Lindeberg 条件, 231, 233, 259
Logistic 回归模型, 235

Marcinkiewicz 不等式, 168
Markov 链, 170
Minimax 准则, 22
Minimax 估计, 90, 91

Newton-Raphson 迭代程序, 227
Neyman-Pearson 基本引理, 108
Neyman-Pearson 基本引理的推广, 114

Pitman 估计, 66, 100
Prohorov 定理, 248, 253

Radon-Nikodym 定理, 4
Rao-Blackwell 定理, 42, 50

Skorohod 表示定理, 249
Slutsky 定理, 249

Tobit 模型, 236

UMA 置信区间和置信限, 146

Wilk-Shapiro 的 W 检验, 211
Winsor 化均值 (Winsorized mean), 230

B

包络函数, Envelope function, 255
边项, Extreme term, 219
变换群的轨道 (orbit), 64
不变检验, 125
不变量和极大不变量, 65, 126

C

参数, 参数空间, 12
参数的集估计, 135
乘积测度, 4
乘积空间, 251
充分统计量, 28
充分性原则, 33
抽样分布, 24
垂直密度表示, 156
次序统计量, 5, 24, 214
次序统计量的分布, 216
次序统计量的极限分布, 218
次序统计量的条件分布, 218
次序样本, 5, 214

D

单边截断族, 52
单调似然比分布族, MLR 族, 110
单指标模型, Single index model, 237
第一类错误和第二类错误, 108
点估计, 48
独立性检验, 208
多个参数情形的 UMVU 估计, 59
多余参数, 120

E

二值响应模型, 236

F

方差一致最小的无偏估计, UMVU 估计, 49
非参数回归模型, 238
非随机化估计, 48
非随机化检验, 107
非线性回归模型, 237
分位函数, Quantile function, 215
分位数, Quantile, 24, 72, 180, 215
分位数变换, Quantile transformation, 217
分位数回归, 232
风险函数, 20

G

概率测度的弱收敛 (依分布收敛), 247
共轭先验分布族, 74

估计的均方误差 (MSE), 48
估计的平均绝对误差, 48
估计的容许性, 94
估计量, 48
广义极大似然估计, GMLE, 85
广义先验分布, 78
广义线性模型, 234

H

后验分布, Posterior distribution, 70
后验风险, Posterior risk, 70
回复水平 (return level), 224
回复周期 (return period), 224

J

极大不等式, Maximal inequalities, 257
极大拟似然估计, 235
极大似然估计, MLE, 170, 235
极小充分统计量, 35
极值分布及其吸引场, 221
假设检验问题的 Neyman-Pearson 提法, 107
检验的功效函数, 107
检验的水平, 107
检验的相似性与 Neyman 结构, 116
检验函数, 107
检验函数的弱紧性, 245
简单假设和复合假设, 107
渐近可测 (asymptotically measurable), 253
渐近胎紧 (asymptotically tight), 253
渐近有效估计, AEE, 164
渐近正态性 (Asymptotic normality), 163
接受域, 否定 (拒绝) 域, 107
截尾和删失, 86
经验 Bayes (Empirical Bayes) 估计, 79
经验分布, 24, 182, 254
经验分布函数, 215, 254
经验分位函数, 215
经验过程, 182, 254
矩估计, Moment estimates, 166
距离空间上的 Borel-可测函数, 250
距离空间上的 Borel σ-域, 250
距离空间上的随机元, 250
决策函数的 Bayes 风险, 69
决策函数列的正则收敛, 77
决策空间, 行动空间, 19
决策问题的不变性, 61

K

可测变换, 1
可测空间, 1
可递变换群, 64
可信水平, 可信系数, 149
空间 $l^{\infty}(T)$, 250
空间 $UC(T,\rho)$, 251
空间 $C[a,b]$, 251
空间 $D[a,b]$, 251
括号数, 255
括号熵, 255

L

连续映照定理, 248, 253
两决策问题, 113
两样本的 Smirnov 检验, 213
列联表 Contingency table, 208

M

密度函数, 11

N

拟合优度检验, Goodness of fit tests, 203
逆映照存在定理, 167

P

平均相合性, 163

Q

齐一性检验, 209
契合性 (Contiguity), 194

强相合性, 163
切尾均值 trimmed mean, 230
区间估计, 135

R

容许性, 22

S

三大类极值分布
    Fréchet 分布, 212
    Weibull 分布, 212
    Gumbel 分布, 222
熵距离, 40
受控分布族, 11
枢轴变量, 137
双边截断族, 53
似然比检验, Likelihood ratio tests, 199
似然方程, 对数似然方程, 170
似然函数 (Likelihood function), 170
随机函数序列的弱收敛, 253
随机化估计, 48
随机化检验, 107
随机化决策函数, 21
随机元, 2
损失函数, 20

T

胎紧 (tight, tightness), 248, 253
条件 Jensen 不等式, 43
条件概率, Conditional probability, 7
条件概率分布, Conditional distribution, 7
条件期望, Conditional expectation, 5
同变估计, Equivariant estimates, 61
同变决策函数, 62
同变性 (Equivariance), 61
同时区间估计, 136
同时置信限, 136
统计假设的检验, 原假设, 对立假设, 106
统计决策函数, 18, 20
统计量, 23
统计模型的均方可微性, 189

W

外概率, 251
外期望, 251
完全类和本质完全类, 22
完全性, 35
无偏估计, 49
无偏检验, Unbiased Test, 116
无信息先验, Noninformative prior, 73, 87

X

先验分布, Prior distribution, 69
线性回归模型, 230
相对紧 (relative compactness), 248
相合性 (弱相合性) Consistency, 163
信仰分布, 153
信仰区间, 153
信仰水平, 信仰系数, 153
信仰推断法 Fiducial Inference, 152

Y

样本, 样本分布, 10
样本分位数, Sample quantile, 215
样本极大值, Sample maxima, 216
样本极小值, Sample minima, 216
样本极值, Sample extreme value, 216, 219
样本空间, 欧氏样本空间, 2
样本量, 11
样本相关系数, 166
样本中位数, Sample median, 215
一揽子引理, Portmanteau Lemma, 248, 252
一列映照到一个随机元的弱收敛, 252
一些常用分布族, 16
一致可积, 250
一致最优不变检验, UMP 不变检验, UMPI 检验, 126

一致最优检验, UMP 检验, 108
一致最优无偏检验, UMP 无偏检验, UMPU 检验, 117
依概率连续, 186
依概率收敛, 247, 252
依概率有界, 248
因子分解定理, 30
有共同支撑的分布族, 12

Z

正则中心项, Regular central term, 219
正则中心项序列的 Bahadur 表示, 219
指数分布族 $\{E(\alpha,\beta)\}$, 13
指数型分布族, 指数族, 12
指数族的满秩条件, 15, 172
置信集, 置信区域, 136
置信区间, 136
置信上限, 置信下限, 136
置信水平, 置信系数, 136
秩统计量 (Rank statistics), 24, 214
中位数, Median, 24, 72, 180, 215
中位数回归, 232
自然参数空间, 13
总体, 总体分布, 10
最高概率密度区域, HPD 区域, 148
最高后验密度区域, HPD 区域, 149
最高信仰密度区域, HFD 区域, 156
最小二乘估计, LS 估计, 180, 230
最小风险同变估计, MRE 估计, 63
最小风险无偏估计, MRU 估计, 50
最小绝对偏差估计, LAD 估计, 180, 232
最优渐近正态估计, BAN 估计, 165