现代数学译丛 3

现代非参数统计

〔美〕L.沃塞曼 著

吴喜之 译

3

科学出版社
www.sciencep.com

现代数学译丛 3

现代非参数统计

〔美〕 L. 沃塞曼 著
吴喜之 译

科学出版社
北京

图字：01-2007-3532 号

内 容 简 介

本书是“All of Nonparametric Statistics”的中译本，源于作者为研究生开设的课程讲义，包括了几乎所有的现代非参数统计的内容. 这种包罗万象的书不但国内没有，在国外也很难找到. 本书主要包括 10 章内容，主要讲述非参数 delta 方法和自助法之类的经验 CDF、覆盖基本的光滑方法和正态均值、利用正交函数的非参数推断、小波和其他的适应方法等.

本书适合统计和计算机科学的高年级大学生、研究生作为教材，也适合统计学、计算机和数据挖掘等方向的研究人员参考.

图书在版编目(CIP)数据

现代非参数统计/(美)L. 沃塞曼著；吴喜之译. —北京：科学出版社，2008
(现代数学译丛；3)
ISBN 978-7-03-021229-0

I. 现… II. ①沃… ②吴… III. 非参数统计 IV. O212.7

中国版本图书馆 CIP 数据核字(2008) 第 027559 号

责任编辑：陈玉琢 / 责任校对：陈玉凤
责任印制：赵德静 / 封面设计：王　浩

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

铭浩彩色印装有限公司印刷

科学出版社发行　各地新华书店经销

*

2008 年 5 月第 一 版　开本：B5(720 × 1000)
2008 年 5 月第一次印刷　印张：14 3/4
印数：1—3 000　字数：271 000

定价：48.00 元

(如有印装质量问题，我社负责调换〈长虹〉)

译 者 的 话

非参数统计可以分为两个范畴, 一个是比较经典的基于秩的, 以检验为主的非参数统计推断, 而另一部分是近二三十年来发展的非参数回归、非参数密度估计、自助法以及小波方法等现代非参数统计方法. 这两者均不对总体分布做较为确定假定, 但除此之外, 这两部分内容在方法上和概念上均没有多少共同点. 这本书就是涉及后者的现代非参数统计的内容.

这本书之所以吸引人, 就在于它包括了几乎所有的现代非参数统计的内容. 这种包罗万象的书不但国内没有, 在国外也很难找到. 这对于读者实在是太方便了. 这也是原书名为"非参数统计大全"(*All of Nonparametric Statistics*) 的原因. 为了包含更多的内容, 本书省略了许多证明细节, 这对于多数研究者来说并不会带来不便. 那些想知道个别数学细节的人, 可以从本书的参考文献中得到满足, 而大多数读者则会满足于本书的全面性和整体性.

吴喜之

2008 年 5 月于中国人民大学统计学院

前 言

目前已经有了许多涉及各种非参数推断的书，如密度估计、非参数回归、自助法及小波方法等. 然而, 很难在一本书中找到所有这些内容. 本教材的宗旨就是为了在一本书中简单扼要地介绍非参数推断的许多现代课题.

本书以统计和计算机科学的硕士或博士生水平的读者为对象. 也适用于想在现代非参数方法方面自学速成的那些在统计学、机器学习和数据挖掘等方向的研究人员. 我的目标是让读者很快熟悉许多领域的基本概念, 而不是纠缠在一个题目上讨论大量的细节. 一方面要覆盖大量的内容, 另一方面要保持本书的精炼. 我决定省略许多证明. 读者可以从本书引用的文献中找到进一步的细节. 当然, 尽管本书标题那么写, 我还是必须在包含什么和省略什么内容方面作出选择. 大体上说, 我决定略去在一章中无法容下的题目. 例如, 不涉及分类或非参数贝叶斯推断.

本书源自我主要为硕士生所开的半学期 (20 小时) 课程的讲义. 教师可能想要求博士生学习更深入的内容并要求他们证明某些定理. 我始终试图遵照我的基本原则, 即绝不给出一个没有置信集的估计量.

本书是方法和理论的混合, 其内容为一些更加注重方法的教材的补充. 这些教材包括 Hastie et al. (2001) 和 Ruppert et al. (2003).

在第 1 章的引言之后, 第 2, 3 章涉及诸如非参数 delta 方法和自助法之类的经验 CDF. 第 4~6 章覆盖基本的光滑方法. 第 7~9 章有较高等的理论内容并且更难些. 第 7 章奠定了第 8, 9 章正交函数法所需要的基础. 第 10 章概述了某些略掉的内容.

我假定读者学过像 Casella and Berger (2002) 或 Wasserman (2004) 的教科书那样的数理统计课程. 特别地, 假定读者熟悉下面的概念：分布函数、依概率收敛、依分布收敛、几乎处处收敛、似然函数、最大似然、置信区间、delta 方法、偏差、均方误差及贝叶斯估计量等. 第 1 章将简单回顾这些背景概念.

数据集和代码能够在下面网址找到.

`www.stat.cmu.edu/~larry/all-of-nonpar.`

我需要做些澄清. 首先, 本书的内容是在"现代非参数统计"的标题之下, 略去诸如秩检验那样的传统方法并不贬低它们的重要性. 其次, 我大量利用大样本方法. 这部分地因为我认为统计大体上在大样本情况下是最成功和有用的, 也部分地因为构造大样本非参数方法常常更加容易. 读者应该意识到, 大样本方法在不够谨慎时

自然会误导.

我谨感谢下面提供了反馈和建议的人士: Larry Brown, Ed George, John Lafferty, Feng Liang, Catherine Loader, Jiayang Sun 及 Rob Tibshirani. 特别要感谢一些提供了非常详细评论的读者: Taeryon Choi, Nils Hjort, Woncheol Jang, Chris Jones, Javier Rojo, David Scott 及一个匿名读者. 还要感谢我的同事 Chris Genovese, 他提出了大量的建议并且为本书的版式提供了 LaTex 宏. 我欠了 John Kimmel 很多, 他一直对我予以支持和帮助, 并且对本书的另类标题不表异议. 最后, 感谢我的夫人 Isabella Verdinelli, 因为她的建议改进了本书, 也因为她的爱心和支持.

L. 沃塞曼

2005 年 7 月于宾夕法尼亚　匹兹堡

目　录

第1章　引　　言

本章将简要地描述将要涉及问题的类型. 然后, 定义某些符号并回顾概率论和统计推断的一些基本概念.

1.1　什么是非参数推断

非参数推断的基本思想是在尽可能少的假定时利用数据对一个未知量作出推断. 通常, 这意味着利用具有无穷维的统计模型. 的确, 对非参数推断的一个更好的名字可能是无穷维推断. 但很难给出非参数推断一个精确的定义, 而且如果一定要给出一个, 将会毫无疑问地遭到反对观点的炮轰.

为了本书的目的, 将把非参数推断这个词组用于旨在保持背景假定尽可能少的现代统计方法的一个集合. 具体地说, 将考虑下面问题:

(1) *估计分布函数*. 给定一个 IID 样本 $X_1,\cdots,X_n \sim F$, 估计 CDF $F(x) = \mathbb{P}(X \leqslant x)$ (第 2 章).

(2) *估计泛函*. 给定一个 IID 样本 $X_1,\cdots,X_n \sim F$, 估计一个泛函 $T(F)$, 如均值 $T(F) = \int x\mathrm{d}F(x)$ (第 2, 3 章).

(3) *密度估计*. 给定一个 IID 样本 $X_1,\cdots,X_n \sim F$, 估计密度 $f(x) = F'(x)$ (第 4, 6, 8 章).

(4) *非参数回归或曲线估计*. 给定 $(X_1,Y_1),\cdots,(X_n,Y_n)$, 估计回归函数 $r(x) = \mathbb{E}(Y|X=x)$ (第 4, 5, 8, 9 章).

(5) *正态均值*. 给定 $Y_i \sim N(\theta_i,\sigma^2)$, $i=1,\cdots,n$, 估计 $\boldsymbol{\theta}=(\theta_1,\cdots,\theta_n)$. 这个看似简单的问题实际上非常复杂, 而且还对大量非参数推断提供了一个统一的偏倚 (第 7 章).

此外, 将在第 7 章讨论某些统一的理论上的原则. 在第 10 章考虑几个不同性质的问题, 如测量误差、逆问题及检验.

最有代表性的假定将是: 分布函数 F(或者密度 f 或回归函数 r) 属于称为**统计模型**(statistical model)的某个大集合 $\mathfrak{F}$. 例如, 在估计密度 f 时, 可能假定

$$f \in \mathfrak{F} = \left\{ g : \int [g''(x)]^2 \mathrm{d}x \leqslant c^2 \right\},$$

它是并不 "太波动" 的密度集合.

1.2　符号和背景知识

这里给出某些有用符号和背景知识的汇总, 参见表 1.1.

表 1.1　某些有用的符号

符号	定义
$x_n = o(a_n)$	$\lim\limits_{n\to\infty} x_n/a_n = 0$
$x_n = O(a_n)$	对大的 n, $\lvert x_n/a_n\rvert$ 有界
$a_n \sim b_n$	当 $n \to \infty$ 时, $a_n/b_n \to 1$
$a_n \asymp b_n$	对大的 n, a_n/b_n 和 b_n/a_n 有界
$X_n \rightsquigarrow X$	依分布收敛
$X_n \xrightarrow{\text{P}} X$	依概率分布
$X_n \xrightarrow{\text{a.s.}} X$	几乎处处收敛
$\widehat{\theta}_n$	参数 θ 的估计
bias	$\mathbb{E}(\widehat{\theta}_n) - \theta$
se	$\sqrt{\mathbb{V}(\widehat{\theta}_n)}$(标准误差)
$\widehat{\text{se}}$	估计的标准误差
MSE	$\mathbb{E}(\widehat{\theta}_n - \theta)^2$(均方误差)
Φ	标准正态随机变量的 CDF
z_α	$\Phi^{-1}(1-\alpha)$

令 $a(x)$ 是 x 的一个函数, 而 F 是一个累积分布函数. 如果 F 是绝对连续的, 令 f 表示其密度. 如果 F 是离散的, 则令 f 为其概率分布函数. a 的均值为

$$\mathbb{E}(a(X)) = \int a(x)\mathrm{d}F(x) \equiv \begin{cases} \displaystyle\int a(x)f(x)\mathrm{d}x, & \text{连续情况,} \\ \displaystyle\sum_j a(x_j)f(x_j), & \text{离散情况.} \end{cases}$$

令 $\mathbb{V} = \mathbb{E}(X - \mathbb{E}(X))^2$ 表示一个随机变量的方差. 如果 $X_1, \cdots, X_n$ 是 n 个观测值, 那么, $\displaystyle\int a(x)\mathrm{d}\widehat{F}_n(x) = n^{-1}\sum_i a(X_i)$, 这里 $\widehat{F}_n$ 是 经验分布 (empirical distribution), 它在每个观测值 x_i 都分配了概率 $1/n$.

对概率论的简单回顾. **样本空间** (sample space)Ω 是一个实验的所有可能结果的集合. Ω 的子集称为**事件** (event). 一个事件类 $\mathcal{A}$ 如果满足下面三个条件则称为一个 σ**域** (σ-field): (i) $\varnothing \in \mathcal{A}$; (ii) $A \in \mathcal{A}$ 意味着 $A^c \in \mathcal{A}$; (iii) $A_1, A_2, \cdots, \in \mathcal{A}$ 意味着 $\bigcup\limits_{i=1}^{\infty} A_i \in \mathcal{A}$. 一个**概率测度** (probability measure)是定义在一个 σ 域 $\mathcal{A}$ 上的函数, 它满足: 对于所有的 $A \in \mathcal{A}$, $\mathbb{P}(A) \geqslant 0$, $\mathbb{P}(\Omega) = 1$, 及如果 $A_1, A_2, \cdots, \in \mathcal{A}$ 是不相交

的, 则

$$\mathbb{P}\left(\bigcup_{i=1}^{\infty} A_i\right) = \sum_{i=1}^{\infty} \mathbb{P}(A_i).$$

三元组合 $(\Omega, \mathcal{A}, \mathbb{P})$ 就称为一个**概率空间** (probability space). 一个**随机变量** (random variable) 是一个映射 $X : \Omega \to \mathbb{R}$, 对于每个实数 x, 都有 $\{\omega \in \Omega : X(\omega) \leqslant x\} \in \mathcal{A}$.

考虑一个随机变量序列 X_n 和随机变量 X. 如果当 $n \to \infty$ 时, 极限

$$\mathbb{P}(X_n \leqslant x) \to \mathbb{P}(X \leqslant x) \tag{1.1}$$

在所有 CDF

$$F(x) = \mathbb{P}(X \leqslant x) \tag{1.2}$$

连续的点 x 成立, 那么称随机变量序列 X_n**依分布收敛**(converge in distribution) 或弱收敛 (converge weekly) 到 X, 记为 $X_n \rightsquigarrow X$. 如果当 $n \to \infty$ 时, 对于每个 $\epsilon > 0$,

$$\mathbb{P}(|X_n - X| > \epsilon) \to 0, \tag{1.3}$$

那么称随机变量序列 X_n**依概率收敛** (converge in probability)到 X, 记为 $X_n \xrightarrow{\mathrm{P}} X$. 如果

$$\mathbb{P}(\lim_{n\to\infty} |X_n - X| = 0) = 1, \tag{1.4}$$

则称随机变量序列 X_n**几乎处处收敛** (converge almost surely)到 X, 记为 $X_n \xrightarrow{\mathrm{a.s.}} X$. 下面的递推关系成立:

$$X_n \xrightarrow{\mathrm{a.s.}} X \Rightarrow X_n \xrightarrow{\mathrm{P}} X \Rightarrow X_n \rightsquigarrow X. \tag{1.5}$$

令 g 为一个连续函数. 按照**连续映射定理** (continuous mapping theorem),

$$\begin{aligned} X_n \rightsquigarrow X &\Rightarrow g(X_n) \rightsquigarrow g(X), \\ X_n \xrightarrow{\mathrm{P}} X &\Rightarrow g(X_n) \xrightarrow{\mathrm{P}} g(X), \\ X_n \xrightarrow{\mathrm{a.s.}} X &\Rightarrow g(X_n) \xrightarrow{\mathrm{a.s.}} g(X). \end{aligned}$$

按照 Slutsky 定理, 如果对于某个常数 c, $X_n \rightsquigarrow X$ 而且 $Y_n \rightsquigarrow c$, 那么, $X_n + Y_n \rightsquigarrow X + c$ 及 $X_n Y_n \rightsquigarrow cX$.

令 $X_1, \cdots, X_n \sim F$ 为 IID. **弱大数定理**(weak law of large numbers) 叙述, 如果 $\mathbb{E}|g(X_1)| < \infty$, 那么, $n^{-1}\sum_{i=1}^{n} g(X_i) \xrightarrow{\mathrm{P}} \mathbb{E}(g(X_1))$. **强大数定理** (strong law of large number)叙述, 如果 $\mathbb{E}|g(X_1)| < \infty$, 那么, $n^{-1}\sum_{i=1}^{n} g(X_i) \xrightarrow{\mathrm{a.s.}} \mathbb{E}(g(X_1))$.

如果随机变量 Z 有密度 $\phi(z)=(2\pi)^{-1/2}\mathrm{e}^{-z^2/2}$, 那么它有标准正态分布, 记为 $Z\sim N(0,1)$. 其 CDF 记为 $\Phi(z)$. 其上 α 分位点记为 z_α. 这样, 如果 $Z\sim N(0,1)$, 则 $\mathbb{P}(Z>z_\alpha)=\alpha$.

如果 $\mathbb{E}(g^2(X_1))<\infty$, **中心极限定理叙述** (central limit theorem),

$$\sqrt{n}(\overline{Y}_n-\mu)\rightsquigarrow N(0,\sigma^2), \tag{1.6}$$

这里, $Y_i=g(X_i),\mu=\mathbb{E}(Y_1),\overline{Y}_n=n^{-1}\sum_{i=1}^{n}Y_i$ 及 $\sigma^2=\mathbb{V}(Y_1)$. 一般来说, 如果

$$\frac{(X_n-\mu)}{\widehat{\sigma}_n}\rightsquigarrow N(0,1),$$

那么, 写

$$X_n\approx N(\mu,\widehat{\sigma}_n^2). \tag{1.7}$$

按照**delta 方法**(delta method), 如果 g 为在 μ 可微的, 并且 $g'(\mu)\neq 0$, 则

$$\sqrt{n}(X_n-\mu)\rightsquigarrow N(0,\sigma^2)\Rightarrow\sqrt{n}[g(X_n)-g(\mu)]\rightsquigarrow N(0,(g'(\mu))^2\sigma^2). \tag{1.8}$$

类似的结果在向量情况也成立. 假定 $\boldsymbol{X}_n$ 为一个随机向量序列, 满足 $\sqrt{n}(\boldsymbol{X}_n-\boldsymbol{\mu})\rightsquigarrow N(0,\boldsymbol{\Sigma})$, 这是具有 0 均值和协方差矩阵 $\boldsymbol{\Sigma}$ 的一个多元正态分布. 令 g 为可微的, 有梯度 ∇_g, 满足 $\nabla_\mu\neq 0$, 这里 ∇_μ 为在 $\boldsymbol{\mu}$ 的 ∇_g. 那么,

$$\sqrt{n}[g(\boldsymbol{X}_n)-g(\boldsymbol{\mu})]\rightsquigarrow N\left(0,(\nabla_\mu^{\mathrm{T}}\boldsymbol{\Sigma}\nabla_\mu\right). \tag{1.9}$$

统计概念. 令 $\mathfrak{F}=\{f(x;\theta):\theta\in\Theta\}$ 为一个满足适当正则条件的参数模型. 基于 IID 观测 $X_1,\cdots,X_n$ 的**似然函数** (likelihood function)为

$$\mathcal{L}_n(\theta)=\prod_{i=1}^{n}f(X_i;\theta),$$

而**对数似然函数** (log-likelihood function)为 $\ell_n(\theta)=\log\mathcal{L}(\theta)$. 最大似然估计, 或 MLE $\widehat{\theta}_n$, 是使似然函数最大的 θ. **得分函数** (score function)为 $s(X;\theta)=\partial\log f(X;\theta)/\partial\theta$. 在适当的正则条件下, 得分函数满足 $\mathbb{E}_\theta(s(X;\theta))=\int s(x;\theta)f(x;\theta)\mathrm{d}x=0$. 此外,

$$\sqrt{n}(\widehat{\theta}_n-\theta)\rightsquigarrow N(0,\tau^2(\theta)),$$

这里 $\tau^2(\theta)=1/I(\theta)$, 而

$$I(\theta)=\mathbb{V}_\theta(s(x;\theta))=\mathbb{E}_\theta(s^2(x;\theta))=-\mathbb{E}_\theta\left(\frac{\partial^2\log f(x;\theta)}{\partial\theta^2}\right)$$

为**Fisher 信息**(Fisher information). 还有,

$$\frac{(\widehat{\theta}_n - \theta)}{\widehat{\mathrm{se}}} \rightsquigarrow N(0,1),$$

这里 $\widehat{\mathrm{se}}^2 = 1/(nI(\widehat{\theta}_n))$. 源于 n 个观测的 Fisher 信息阵 I_n 满足 $I_n(\theta) = nI(\theta)$; 因此, 也记 $\widehat{\mathrm{se}}^2 = 1/(I_n(\widehat{\theta}_n))$.

一个估计 $\widehat{\theta}_n$ 的偏倚为 $\mathbb{E}(\widehat{\theta}) - \theta$ 而且均方误差 MSE 为 MSE$= \mathbb{E}(\widehat{\theta} - \theta)^2$. 对估计 $\widehat{\theta}_n$ 的 MSE 的**偏倚–方差分解**(bias-variance decomposition)为

$$\mathrm{MSE} = \mathrm{bias}^2(\widehat{\theta}_n) + \mathbb{V}(\widehat{\theta}_n). \tag{1.10}$$

1.3 置 信 集

非参数推断多是为了发现对某感兴趣的量 θ 的估计 $\widehat{\theta}_n$. 例如, 这里的 θ 可能是一个均值, 一个密度或者一个回归函数. 但是仍然想要提供关于这些量的置信集. 正如将要解释的, 有不同形式的置信集.

令 $\mathfrak{F}$ 为分布函数 F 的一个类, 而 θ 为某个感兴趣的量. 这样, θ 可能是 F 本身, 或者 F', 或者是 F 的均值等等. 令 C_n 为 θ 可能取值的集合, 它依赖于数据 $X_1, \cdots, X_n$. 为了强调概率的陈述依赖于背景中的 F, 有时记它为 $\mathbb{P}_F$.

1.11 定义 · 如果

$$\inf_{F\in\mathfrak{F}} \mathbb{P}_F(\theta \in C_n) \geqslant 1-\alpha, \quad \text{对所有 } n \text{ 成立}, \tag{1.12}$$

那么 C_n 为一个有穷样本 $1-\alpha$**置信集**(finite sample $1-\alpha$ confidence set). 如果

$$\liminf_{n\to\infty} \inf_{F\in\mathfrak{F}} \mathbb{P}_F(\theta \in C_n) \geqslant 1-\alpha, \tag{1.13}$$

那么 C_n 为一个一致渐近 $1-\alpha$**置信集**(uniform asymptotic $1-\alpha$ confidence set). 如果

$$\liminf_{n\to\infty} \mathbb{P}_F(\theta \in C_n) \geqslant 1-\alpha, \quad \text{对每个 } F\in\mathfrak{F} \text{ 成立}, \tag{1.14}$$

那么 C_n 为一个逐点渐近 $1-\alpha$**置信集**(pointwise asymptotic $1-\alpha$ confidence set).

如果 $||\cdot||$ 表示某种范数, 而 $\widehat{f}_n$ 为 f 的一个估计值, 那么, 关于 f 的一个**置信球** (confidence ball)为有下面形式的一个置信集:

$$C_n = \{f \in \mathfrak{F}:\ ||f - \widehat{f}_n|| \leqslant s_n\}, \tag{1.15}$$

这里 s_n 可能依赖于数据. 假定 f 定义在集合 $\mathcal{X}$ 上. 对于函数对 (ℓ, u), 如果

$$\inf_{f \in \mathfrak{F}} \mathbb{P}_F(\ell(x) \leqslant f(x) \leqslant u(x), \quad \text{对所有 } x \in \mathcal{X}) \geqslant 1-\alpha, \tag{1.16}$$

那么 (ℓ, u) 称为一个$1-\alpha$ **置信带** ($1-\alpha$ confidence band)或**置信包络**(confidence envelope). 置信球和置信带可能是上面所说的有穷样本, 逐点渐近和一致渐近的. 当估计一个实数值的量而不是一个函数时, C_n 恰好是一个区间, 则称 C_n 为置信区间.

理想地, 希望找到有穷样本置信集. 当这是不可能时, 则试图构造一致渐近置信集. 而最后的选择是求逐点渐近置信集. 如果 C_n 为一个一致渐近置信集, 那么下面的说法为真: 对任何 $\delta > 0$, 存在一个 $n(\delta)$, 使得对所有 $n > n(\delta)$, C_n 的收敛至少为 $1-\alpha-\delta$. 对于逐点渐近置信集, 可能不存在一个有穷的 $n(\delta)$. 这时, 使得置信集收敛接近 $1-\alpha$ 的样本量将依赖于 (未知的)f.

1.17 例 令 $X_1, \cdots, X_n \sim \text{Bernoulli}(p)$. 关于 p 的一个逐点渐近 $1-\alpha$ 置信区间为

$$\widehat{p}_n \pm z_{\alpha/2}\sqrt{\frac{\widehat{p}_n(1-\widehat{p}_n)}{n}}, \tag{1.18}$$

这里 $\widehat{p}_n = n^{-1}\sum_{i=1}^{n} X_i$. 根据 Hoeffding 不等式 (1.24), 一个有穷样本置信区间为

$$\widehat{p}_n \pm \sqrt{\frac{1}{2n}\log\left(\frac{2}{\alpha}\right)}. \tag{1.19}$$

■

1.20 例(参数模型) 令

$$\mathfrak{F} = \{f(x);\theta) : \ \theta \in \Theta\}$$

为有纯量参数 θ 的参数模型, 而 $\widehat{\theta}_n$ 为最大似然估计, 它是使得似然函数

$$\mathcal{L}_n(\theta) = \prod_{i=1}^{n} f(X_i;\theta)$$

最大的 θ 值. 回顾在适当的正则假定下,

$$\widehat{\theta}_n \approx N(\theta, \widehat{\text{se}}^2),$$

这里,

$$\widehat{\text{se}} = [I_n(\widehat{\theta}_n)]^{-1/2}$$

为被估计的 $\widehat{\theta}_n$ 的标准误差, 而 $I_n(\theta)$ 为 Fisher 信息. 这样,

$$\widehat{\theta}_n \pm z_{\alpha/2}\widehat{\text{se}}$$

为一个逐点渐近置信区间. 如果 $\tau = g(\theta)$, 能够利用 delta 方法得到一个 τ 的渐近置信区间. τ 的 MLE 为 $\widehat{\tau}_n = g(\widehat{\theta}_n)$. 对于 τ, 被估计的标准误差为 $\widehat{\text{se}}(\widehat{\tau}_n) = \widehat{\text{se}}(\widehat{\theta}_n)|g'(\widehat{\theta}_n)|$. τ 的条件置信区间为

$$\widehat{\tau}_n \pm z_{\alpha/2}\widehat{\text{se}}(\widehat{\tau}_n) = \widehat{\tau}_n \pm z_{\alpha/2}\widehat{\text{se}}(\widehat{\theta}_n)|g'(\widehat{\theta}_n)|.$$

这又是一个典型的逐点渐近置信区间. ■

1.4 有用的不等式

本书不时需要用某些不等式. 为了查阅方便的目的, 一些不等式在下面列出:

Markov 不等式. 令 X 为非负随机变量, 并假定 $\mathbb{E}(X)$ 存在. 对于任何 $t > 0$,

$$\mathbb{P}(X > t) \leqslant \frac{\mathbb{E}(X)}{t}. \tag{1.21}$$

Chebyshev 不等式. 令 $\mu = \mathbb{E}(X)$, 而且 $\sigma^2 = \mathbb{V}(X)$. 那么,

$$\mathbb{P}(|X - \mu| \geqslant t) \leqslant \frac{\sigma^2}{t^2}. \tag{1.22}$$

Hoeffding 不等式. 令 $Y_1, \cdots, Y_n$ 为独立观测, 满足 $\mathbb{E}(Y_i) = 0$ 及 $a_i \leqslant Y_i \leqslant b_i$. 令 $\epsilon > 0$. 那么, 对任何 $t > 0$,

$$\mathbb{P}\left(\sum_{i=1}^{n} Y_i \geqslant \epsilon\right) \leqslant \mathrm{e}^{-t\epsilon} \prod_{i=1}^{n} \mathrm{e}^{t^2(b_i - a_i)^2/8} \tag{1.23}$$

为 Bernoulli 随机变量的 Hoeffding 不等式. 令 $X_1, \cdots, X_n \sim \text{Bernoulli}(p)$. 那么, 对任何 $\epsilon > 0$,

$$\mathbb{P}(|\overline{X}_n - p| > \epsilon) \leqslant 2\mathrm{e}^{-2n\epsilon^2}, \tag{1.24}$$

这里, $\overline{X}_n = n^{-1}\sum_{i=1}^{n} X_i$.

Mill 不等式. 如果 $Z \sim N(0,1)$, 那么, 对任何 $t > 0$,

$$\mathbb{P}(|Z| > t) \leqslant \frac{2\phi(t)}{t}, \tag{1.25}$$

这里, ϕ 是标准正态密度. 事实上, 对于任何 $t > 0$,

$$\left(\frac{1}{t} - \frac{1}{t^3}\right)\phi(t) < \mathbb{P}(Z > t) < \frac{1}{t}\phi(t) \tag{1.26}$$

及

$$P(Z > t) < \frac{1}{2}\mathrm{e}^{-t^2/2}. \tag{1.27}$$