北京工业大学研究生创新教育系列教材

应用非参数统计

薛留根　编著

科学出版社

北京工业大学研究生创新教育系列教材

应用非参数统计

薛留根　编著

科学出版社

北京

内 容 简 介

本书介绍非参数统计的基本概念和方法, 其内容包括: 预备知识, U统计量, 基于二项分布的检验, 列联分析, 秩检验, 检验的功效与渐近相对效率, 概率密度估计, 非参数回归. 每一章内容都着重阐述非参数统计推断的一般处理技术和原则, 并给出一些典型例子. 各章后面的习题侧重于应用. 本书的特点是侧重于介绍非参数统计在各应用领域中的常用方法, 尽可能简化公式推导并淡化理论证明. 此外, 本书有选择地安排一些模拟计算和实际数据分析, 其主要程序放在附录 A 中.

读者只需具有高等数学和概率统计的基本知识即可读懂本书的主要内容. 本书可以作为大学高年级学生或硕士研究生的教材, 也可以作为实际工作者自学的参考书.

图书在版编目(CIP)数据

应用非参数统计/薛留根编著. —北京: 科学出版社, 2013.7

北京工业大学研究生创新教育系列教材

ISBN 978-7-03-038153-8

Ⅰ. ①应… Ⅱ. ①薛… Ⅲ. ①非参数统计-研究生-教材 Ⅳ. ①O212.7

中国版本图书馆 CIP 数据核字 (2013) 第 156453 号

责任编辑: 陈玉琢 / 责任校对: 邹慧卿

责任印制: 钱玉芬 / 封面设计: 陈 敬

科学出版社 出版

北京东黄城根北街 16 号

邮政编码: 100717

http://www.sciencep.com

骏走印刷厂印刷

科学出版社发行 各地新华书店经销

*

2013 年 7 月第 一 版 开本: B5(720×1000)

2013 年 7 月第一次印刷 印张: 14

字数: 271 000

定价: 68.00 元

(如有印装质量问题, 我社负责调换)

前　言

本书是为概率统计专业的学生及相关专业的学生和统计工作者编写的教科书. 阅读本书的读者只需具有高等数学和概率统计的基本知识. 读完本书即可进入非参数统计各相关领域的学习. 本书可以作为大学高年级学生或硕士研究生的教材, 也可以作为实际工作者自学或查阅非参数统计方法的参考书.

全书共分八章, 依次为预备知识, U 统计量, 基于二项分布的检验, 列联分析, 秩检验, 检验的功效与渐近相对效率, 概率密度估计, 非参数回归. 本书着重阐述非参数统计推断的一般处理技术和原则, 并给出一些典型例子. 各章后面的习题侧重于应用.

本书的前身是一本讲义, 作者曾在北京工业大学概率统计学科部作为研究生的 "非参数统计"课程的教材使用了多年. 该讲议虽经过作者多次修改, 但总感不足. 这次趁出版之机, 对本书的一些章节作了较大的修改, 充实了一些新的内容, 在叙述上进行了加工, 尽量使内容既有新意, 又易于理解. 书中丰富的例子着力说明非参数统计的方法和应用, 配置的习题能够让读者得到各种基本训练.

非参数统计的一个特点是它的使用面广, 因为它讨论的模型中分布族没有通过有限个实参数去刻画, 模型使用的范围更大. 因此, 该统计学分支在经济、金融、生物、医学等领域有着广泛的应用. 非参数统计的另一个特点是大样本方法占重要位置. 可以说, 绝大多数常用的非参数统计方法都是基于有关统计量的某种极限性质. 因此, 某些定理的论证很烦琐, 初学者往往感到困难. 考虑到非参数统计这两个特点, 本书在取材与写作上作了三点努力: 一是侧重于介绍非参数统计在各应用领域中的基本方法, 理论的推导和证明尽可能简化; 二是有选择地安排一些模拟计算和应用实例, 强调统计学与计算机结合; 三是在语言叙述上力求简明易懂、严谨系统, 便于阅读和自学.

本书的编写与出版得到了科学出版社陈玉琢编辑的鼓励和关心, 得到了北京工业大学研究生课程建设项目和北京市优秀博士学位论文指导教师科技项目 (编辑:20111000503) 的资助, 并得到了冯三营、刘娟芳和张景华等同志的帮助, 作者谨在此一并表示衷心感谢.

虽然本书在正式出版之前曾作为教材使用了多年, 但由于作者水平有限, 书中不妥之处在所难免, 欢迎国内同行及广大读者不吝指正.

薛留根

2013 年 4 月

目　录

第 1 章　预 备 知 识

本章主要介绍一些预备知识, 其内容包括: 非参数统计概述、数据类型、检验的 p 值、次序统计量及其分布、分位数的估计等.

1.1　非参数统计概述

非参数统计是统计学的一个重要分支. 在学习这门课程之前, 首先要明白什么是 "非参数统计", 了解这个分支的一些基本特点, 从而可以对它有初步的认识, 对学习这门课程产生兴趣.

在统计学中, 统计推断的两个最基本的形式是: 参数估计和假设检验, 其大部分内容是和正态理论相关的, 人们称之为参数统计. 在参数统计中, 总体的分布形式或分布族往往是给定的, 而诸如均值和方差的参数是未知的. 人们的任务就是对这些参数进行估计或检验. 当假定分布成立时, 其推断有较高的精度. 然而, 在实际问题中, 对总体分布的假定并不是总成立, 也就是说, 有时数据并不是来自所假定分布的总体. 因此, 在假定的总体分布下进行推断, 其结果可能会背离实际. 于是, 人们希望在不假定总体分布的情况下, 尽量从数据本身获得所需要的信息. 这就是非参数统计的初衷. 看下面的例子.

例 1.1.1 (概率密度估计)　设随机变量 X 有概率密度函数 $f(x)$, 它属于某个确定的密度族 $\mathcal{F}$. 令 $X_1, \cdots, X_n$ 为 X 的样本, 要通过样本来估计 $f(x)$.

如果 $\mathcal{F}$ 的形式已知, 如正态分布族 $\{N(\mu, \sigma^2), -\infty < \mu < \infty, \sigma^2 > 0\}$, 则只需对分布中的参数 μ 和 σ^2 作出估计, 就可得到密度 $f(x)$ 的估计, 这是一个参数统计问题. 我们可以利用极大似然估计法来估计 μ 和 σ^2.

如果对 $\mathcal{F}$ 只施加一般性的假定, 如 $f(x)$ 对称, 且具有连续的二阶导数等, 则这是一个非参数统计问题. 我们可以利用多种方法对非参数密度函数 $f(x)$ 进行估计, 例如, 核估计法, 最近邻估计法, 小波估计法等. 这些估计方法已成为现代非参数统计的重要内容.

例 1.1.2 (一元回归)　设随机变量 Y 与 X 之间存在着某种相关关系, 这里 X 可以是控制或可以精确观察的变量. 如果在 $X = x$ 的条件下, Y 的数学期望 $E(Y|X = x)$ 存在, 记为 $m(x)$, 则称 $m(x)$ 为 Y 关于 X 的回归函数. 设 $(X_1, Y_1) \cdots$, (X_n, Y_n) 为 (X, Y) 的样本, 要通过样本来估计 $m(x)$.

在一元线性回归模型中, 假定 $m(x)$ 为 x 的线性函数, 即 $m(x)=a+bx$, 且在给定 $X=x$ 的条件下, Y 的分布为正态 $N(a+bx,\sigma^2)$. 这个模型完全由三个实参数 a, b 和 σ^2 所刻画, 而要估计的回归函数 $m(x)$, 实际上只依赖于参数 a 和 b, 因而它是一个典型的参数统计问题. 我们可以利用最小二乘法对 a 和 b 进行估计.

然而, 如果对 Y 的分布不作任何假定, 或只作些一般性假定 (如 Y 的方差有限), 则问题就成为非参数性的, 称为 "非参数回归". 我们可以利用多种方法对非参数回归函数 $m(x)$ 进行估计, 例如, 核估计法, 最近邻估计法, 局部多项式估计法, 小波估计法等. 这些估计方法是现代非参数统计的重要组成部分.

综上所述, 我们可提出下面的定义: 如果一个统计问题的模型所涉及的分布族不能用有限个实参数去刻画, 则该问题称为非参数统计问题. 非参数统计是统计学研究非参数统计问题的一个分支学科.

非参数方法是处理与分布无关的问题的方法. 所谓 "与分布无关", 意味着它的推断方法不假定总体服从确定的分布, 并不是脱离总体的分布. 与参数方法相比, 非参数方法具有下面的特点.

(1) 具有广泛的适用性. 非参数方法不假定具体的总体分布, 从而它适用于来自任何总体分布未知的数据, 能用来描述更多的问题, 故适用面广. 由于非参数方法没有利用关于总体分布的信息, 因此就是在对总体分布没有任何了解的情况下, 它也能获得可靠的结论. 在这一点上, 非参数方法优于参数方法. 然而, 在总体的分布族已知的情况下, 它没有像极大似然估计那样充分利用总体分布的信息, 于是所得出的结论就不如参数方法那样精确, 一般来说效率偏低. 参数方法往往对设定的模型有更大的针对性: 一旦模型改变, 方法也就随之改变. 非参数方法则不然, 由于它对模型的限定少, 以致人们只能用很一般的方式去使用样本中的信息来进行统计推断.

(2) 具有稳健性. 稳健性 (robustness) 反映统计方法这样一种性质: 当真实模型与设定模型的偏离不大时, 这种统计方法仍能保持良好的性质, 至少不至于变得很差. 由于非参数方法对总体分布的限制相对较少, 因此就具有稳健性. 而参数方法是建立在分布已知的基础上, 当总体分布发生改变时, 其推断的正确性就大打折扣, 甚至可能产生错误的结论.

(3) 以大样本理论为主导. 大样本理论在非参数统计中起着重要作用. 可以说, 绝大多数常用的非参数方法都是基于有关统计量的某种极限性质. 非参数统计更多地依赖于大样本方法这一特点, 可以从其模型的广泛性上来理解: 统计量的分布依赖于总体的分布. 如果我们对总体的分布了解很少, 则就难以得出有关统计量的确切分布. 而很多小样本方法是基于这种确切分布的. 例如, 在方差 σ^2 未知的条件下去推断正态总体的期望 μ, 人们就用样本方差 S^2 去代替 σ^2, 然后构造出统计量 $T=\sqrt{n}(\overline{X}-\mu)/S$. 由于当 $n\to\infty$ 时, T 依分布收敛于标准正态分布 $N(0,1)$, 因

此这是一个大样本方法. 但如果总体服从正态分布, 则由 Fisher 基本定理知: T 服从自由度为 $n-1$ 的 t 分布. 因此, 关于 μ 的推断可建立在这个确切分布的基础上, 这就成为一种小样本方法.

1.2 数 据 类 型

在对某个总体进行统计推断时, 首先要从该总体中抽取样本, 然后利用样本构造出统计量, 由此就可以解决估计和检验问题. 数据是样本的观测值, 是样本的实现. 统计工作的主要内容是数据收集和数据处理, 其中数据处理是统计的核心内容, 它是将数据转化为有用信息的过程. 在科学实验和生产实践中, 人们遇到各种各样的数据, 这就为统计分析提供了保障. 然而, 为正确地处理和分析数据, 就必须先了解数据, 这样才能有针对性地选用统计分析方法. 在统计学中，统计数据主要可分为四种类型，分别是定类数据, 定序数据, 定距数据, 定比数据. 定类数据和定序数据称为定性数据; 定距数据和定比数据称为定量数据. 下面我们对这四种类型的数据分别加以介绍.

(1) 定类数据. 某项指标的观测值不是数, 而是事物的属性. 有时, 为了识别不同的类别, 也可以用特定的数字和符号表示某类事物. 例如, 人的性别 (男, 女), 职业 (教师, 医生, 工人), 物体的颜色、样式等, 它们的异同是按照事物的某些特征来划分和辨别. 人们常用数表示属性的分类, 如用数 “1” 和 “0” 分别表示 “男” 和 “女”, 这仅仅是人们赋予的识别代码, 并不说明事物的数量; 它不能进行算术运算, 也没有大小关系, 而只能进行 “=” 或 “$\neq$” 的逻辑运算. 定类数据的描述性统计量有频数、众数等.

(2) 定序数据. 事物的属性具有顺序关系. 为方便起见, 有时也用数字表示. 例如, 家庭经济状况分为高收入、中等收入、低收入三类, 可分别用 3, 2, 1 表示. 这些数只起一个顺序作用, 不能作算术运算, 即这里的 “$3-2$” 是没有意义的. 也就是说, “高收入” 比 “中等收入” 经济状况好, 但 “好多少” 不能计算, 只能比较类别之间的次序关系. 定序数据可以进行 “=” “$\neq$” “$>$” “$<$” 的运算. 描述定序数据集中趋势的最适合统计量是中位数, 反映离散程度的统计量是分位数.

(3) 定距数据. 它说明的是事物的数量特征, 能够用数值表示. 例如, 学生的考试成绩, 某种商品的销售数, 班级的学生数等. 定距数据没有绝对的零点, 如某个学生的考试成绩是 0 分, 这并不表示该生没有这门课的知识. 定距数据不但可以进行 “=” “$\neq$” “$>$” “$<$” 的运算, 而且可以进行 “$+$” 和 “$-$” 的运算. 反映定距数据集中趋势的统计量是均值、中位数、众数, 反映离散程度的统计量是方差、标准差等.

(4) 定比数据. 它说明的是事物的数量特征, 能够用数值表示, 并且有绝对的零点. 例如, 产品的使用寿命, 人的身高、体重, 物体的长度、直径、质量等. 定比数据

不但可以进行 “=” “≠” “>” “<” “+” “−” 的运算, 而且可以进行 “×” 和 “÷” 的运算. 反映定比数据集中趋势和离散程度的描述性统计量不仅有均值、中位数、众数、方差、标准差, 还有变异系数等.

从上述介绍可知, 定性数据描述事物的性质, 其 0 只有相对意义; 定量数据描述事物的数量, 其 0 具有实际意义. 定类数据是最低级别的数据, 定比数据是最高级别的数据, 中间两个级别依次为定序数据和定距数据. 数据的级别越高, 所包含的运算性质就越多.

参数方法所分析的数据主要是定量数据. 非参数方法不但可以用来分析定量数据, 而且还可以用来分析定性数据. 例如, 利用问卷调查资料分析用户对几种商品的喜爱程度是否相等; 利用民意测验分析职工对公司的几种改革方案的支持率是否有差异等. 这方面的研究是参数方法做不到的, 只能应用非参数方法. 这一点又说明了非参数方法应用面广.

当手中有了数据集后, 首先要对它有一个直观的认识. 在数据来自一个总体时, 需要看它的大致分布形状. 利用直方图和 Q-Q 图可以做到这一点. 直方图可以用来看该分布是否呈现出对称性, 是否有很长的尾部. Q-Q 图是按升幂重新排列的样本观测值和标准正态分布的分位数 (通常用 $\Phi^{-1}((i-3/8)/(n+1/4))$ 来作散点图. 如果原来的样本来自正态分布, 则该图应该大致成一条直线; 否则, 它将在一端或两端有摆动, 说明其总体分布与正态分布有差别. 调用统计软件中的函数就可以作出直方图和 Q-Q 图. 如 R 语言中作直方图的函数是 $\mathrm{hist}(x)$, 作 Q-Q 图的函数是 $\mathrm{qqnorm}(x)$, 其中括号中的 x 为数据变量.

1.3　检验的 p 值

给定原假设 H_0 和备择假设 H_1, 并记为检验问题 (H_0, H_1). 为解该检验问题, 首先需要构造检验统计量 T. 然后利用 T 得到检验的拒绝域 W. 最后作出判断: 在 T 的观测值落入 W 时, 就拒绝原假设 H_0, 认为备择假设 H_1 成立; 在 T 的观测值没有落入 W 时, 就不能拒绝原假设 H_0, 只能认为 H_0 成立. 这就是所谓的检验法. 如果引入检验的 p 值, 那么就可以用 p 值对检验作出决定. 检验的 p 值定义如下.

定义 1.3.1　检验的 p 值是在已知观测下拒绝原假设的最小显著性水平. 如果用 t_{obs} 表示检验统计量 T 的观测值, 则左边检验的 p 值是 $P\{T \leqslant t_{\mathrm{obs}}\}$, 右边检验的 p 值是 $P\{T \geqslant t_{\mathrm{obs}}\}$, 双边检验的 p 值是 $P\{T \leqslant t_{\mathrm{obs}}\}$ 和 $P\{T \geqslant t_{\mathrm{obs}}\}$ 中较小者的 2 倍.

严格地讲, 如果 T 的零分布是离散的, 且拒绝域左边和右边的概率不相等, 那么很难构造两边概率相等而精确的显著性水平. 这就与前面的定义不一致. 但为了

避免定义的歧义性, 我们在后面仍认为双边检验的 p 值是观测值落在零分布单边概率的两倍.

从定义 1.3.1 可知, 在 p 值很小时, 说明统计量上的实现在原假设下是小概率事件. 此时, 如果拒绝原假设, 则犯第一类错误 (弃真错误) 的概率也很小, 它等于 p 值. 反之, 如果 p 值很大, 则拒绝原假设所犯第一类错误的概率也大. 因此, 不能拒绝原假设. p 值的具体计算依赖于原假设、统计量的分布及其观测值. 很多统计软件 (包括 R) 对一些常用的假设检验方法都提供了检验的 p 值. 然而, 在现代统计方法研究中, 现有的统计软件中没有直接计算 p 值的函数, 人们只能采用 Monte-Carlo 方法编写程序来计算 p 值.

在实践中, 人们并不事先指定显著性水平, 而是很方便地利用 p 值进行决断. 对于任意大于 p 值的显著性水平, 人们可以拒绝原假设, 但不能在任何小于它的水平下拒绝原假设. 设 p 为计算得到的 p 值. 给定显著性水平 α, 如果 $\alpha \geqslant p$, 则拒绝 H_0, 否则接受原假设.

1.4 次序统计量及其分布

次序统计量在近代统计推断中起着重要作用, 这是由于次序统计量有一些性质不依赖于总体的分布, 并且计算量很小, 使用起来较方便. 因此, 在质量管理、可靠性等方面得到广泛的应用.

定义 1.4.1 设样本 $X_1, \cdots, X_n$ 独立同分布, 把诸 X_i 从小到大按次序排列为

$$X_{(1)} \leqslant X_{(2)} \leqslant \cdots \leqslant X_{(n)},$$

则称 $X_{(1)}, \cdots, X_{(n)}$ 为原样本 $X_1, \cdots, X_n$ 的次序统计量. 称 $X_{(i)}$ 为第 i 个次序统计量. $X_{(1)}$ 为样本的极小值, $X_{(n)}$ 为样本的极大值, 这两者有时通称为“极值”.

设总体 X 的分布函数为 $F(x)$, 且具有连续密度函数 $f(x)$, 则次序统计量 $X_{(r)}$ 的密度函数为

$$f_r(x) = \frac{n!}{(r-1)!(n-r)!}[F(x)]^{r-1}[1-F(x)]^{n-r}f(x).$$

次序统计量 $X_{(r)}$ 与 $X_{(s)}$ 的联合密度函数为

$$\begin{aligned} f_{rs}(x,y) = {} & \frac{n!}{(r-1)!(s-r-1)!(n-s)!}[F(x)]^{r-1}[F(y)-F(x)]^{s-r-1} \\ & \times[1-F(y)]^{n-s}f(x)f(y), \quad x<y. \end{aligned}$$

用类似方法可求得任意三个或更多个次序统计量的联合密度函数. 特别地, $X_{(1)}, \cdots, X_{(n)}$ 的联合密度函数为

$$f(y_1, \cdots, y_n) = n!f(y_1)\cdots f(y_n), \quad y_1 < y_2 < \cdots < y_n.$$

利用上述结果可以推导出一些次序统计量的函数的分布. 例如, 极差 $R = X_{(n)} - X_{(1)}$ 的分布函数为

$$F_R(r) = n\int_{-\infty}^{\infty}[F(x+r) - F(x)]^{n-1}f(x)\mathrm{d}x.$$

1.5　分位数的估计

总体分布的分位数和分位数的函数的估计是非参数估计的基本内容. 这类估计一般不假定总体分布的具体形式. 估计涉及的基本统计量是样本的次序统计量和经验分布函数.

1.5.1　分位数的点估计

设总体分布函数为 $F(x)$. 所谓 $F(x)$ 的 p 分位数 ξ_p 是满足下述条件的一个数:

$$F(\xi_p) \geqslant p, \quad F(\xi_p - 0) \leqslant p, \quad 0 < p < 1,$$

其中 $F(x) = P(X \leqslant x)$, 它是右连续的.

这样定义的 p 分位数不唯一. 易证: 如果分布的 p 分位数不唯一, 则它充满一个有界闭区间.

为了解决唯一性问题, 统计学家又把总体的 p 分位数定义为

$$\xi_p = \inf\{x : F(x) \geqslant p\}, \quad p \in (0,1).$$

当 $p = 1/2$ 时, $\xi_{1/2}$ 为分布的中位数.

估计 ξ_p 的问题是常见的估计问题. 对参数分布族而言, p 分位数常可通过参数表出, 因而估计参数后, 可获得相应的 p 分位数的估计. 但在对总体分布形式未知的情况下, 用样本次序统计量可构成分位数的非参数估计, 即用样本的 p 分位数作为总体分布的 p 分位数的估计. 由此可给出下面的定义.

定义 1.5.1　设 $X_1, \cdots, X_n$ 是来自总体 $F(x)$ 的独立同分布样本, 其经验分布函数记为 $F_n(x) = n^{-1}\sum\limits_{i=1}^{n} I(X_i \leqslant x)$, 则称

$$\hat{\xi}_{n,p} = \inf\{x : F_n(x) \geqslant p\}$$

为样本的 p 分位数.

对于 $\hat{\xi}_{n,p}$, 有下面两个渐近性质.

定理 1.5.1　设总体分布 $F(x)$ 的密度函数 $f(x)$ 在 ξ_p 处连续, 且 $f(\xi_p) > 0$, 则样本分位数 $\hat{\xi}_{n,p}$ 有渐近正态分布 $N\big(\xi_p, p(1-p)/[nf^2(\xi_p)]\big)$.

定理 1.5.2 对 $0<p<1$, ξ_p 是满足 $F(x)\geqslant p$, $F(x-0)\leqslant p$ 的总体 $F(x)$ 的 p 分位数. 如果 ξ_p 是唯一的, 则当 $n\to\infty$ 时, $\hat{\xi}_{n,p}\to\xi_p$, a.s.

上述两个定理的证明可以在有关书籍中找到, 这里省略其证明.

例 1.5.1 设总体分布为均匀分布 $U(0,1)$,

$$F(x)=x,\ 0\leqslant x\leqslant 1,\quad f(x)=1,\quad 0\leqslant x\leqslant 1,$$

$\xi_{1/2}=\dfrac{1}{2}$, $f(x)$ 在此点连续, 故样本中位数 $\hat{\xi}_{n,1/2}$ 具有渐近分布 $N\left(\dfrac{1}{2},\dfrac{1}{4n}\right)$.

1.5.2 分位数的区间估计

1. 大样本区间估计

在大样本情形下, 我们可以利用样本 p 分位数 $\hat{\xi}_{n,p}$ 的渐近正态性构造置信区间. 给定置信水平 $\alpha>0$, 用 $z_{1-\alpha/2}$ 表示满足 $\Phi(z_{1-\alpha/2})=1-\alpha/2$ 的数, 它是标准正态分布的 $1-\alpha/2$ 分位数. 由定理 1.5.1 知

$$\lim_{n\to\infty}P\left\{|\hat{\xi}_{n,p}-\xi_p|\leqslant\frac{z_{1-\alpha/2}\sqrt{p(1-p)}}{\sqrt{n}f(\xi_p)}\right\}=1-\alpha.$$

上式尚不能直接用于区间估计, 因为其中 $f(\cdot)$ 与 ξ_p 皆未知. ξ_p 可用 $\hat{\xi}_{n,p}$ 估计, 至于 $f(\cdot)$, 需用概率密度的非参数估计法估计之. 以 $\hat{f}_n(\cdot)$ 记 $f(\cdot)$ 的一个估计, 如果 $\hat{f}_n(\cdot)$ 有相合性, 则利用上式有

$$\lim_{n\to\infty}P\left\{|\hat{\xi}_{n,p}-\xi_p|\leqslant\frac{z_{1-\alpha/2}\sqrt{p(1-p)}}{\sqrt{n}\hat{f}_n(\hat{\xi}_{n,p})}\right\}=1-\alpha.$$

上式表明, $\hat{\xi}_{n,p}\pm z_{1-\alpha/2}\sqrt{p(1-p)}/[\sqrt{n}\hat{f}_n(\hat{\xi}_{n,p})]$ 是 ξ_p 的一个区间估计, 其渐近置信水平为 $1-\alpha$, 这个估计只有在样本容量 n 相当大时才有用, 因为 n 太小时, 概率密度 $f(\cdot)$ 不易估计准确. 对这种情况, 可使用下面所讲的小样本区间估计.

2. 小样本区间估计

设 $X_1,\cdots,X_n$ 是来自连续分布 $F(x)$ 的一个样本. $X_{(1)}\leqslant\cdots\leqslant X_{(n)}$ 为样本次序统计量. 下面求 p 分位数 ξ_p 的形如 $[X_{(r)},X_{(s)}]$ 的置信区间, 即求最大整数 r 和最小整数 s, 使得

$$P\left\{X_{(r)}\leqslant\xi_p\leqslant X_{(s)}\right\}\geqslant 1-\alpha. \tag{1.5.1}$$

为此, 记 $Y=\sum_{i=1}^{n}I(X_i\leqslant\xi_p)$, 显然 Y 服从二项分布 $B(n,p)$, 其中 $p=P\{X_i\leqslant\xi_p\}$.

注意到事件 $\{X_{(r)} \leqslant \xi_p \leqslant X_{(s)}\}$ 等价于事件 "样本 $X_1, \cdots, X_n$ 中小于等于 ξ_p 的个数至少为 r 且至多为 s", 即等价于事件 $\{r \leqslant Y \leqslant s\}$. 因此,

$$\begin{aligned} & P\left\{X_{(r)} \leqslant \xi_p \leqslant X_{(s)}\right\} \\ = & P\{r \leqslant Y \leqslant s\} = P\{Y \leqslant s\} - P\{Y < r\} \\ = & \sum_{i=0}^{s} \binom{n}{i} p^i (1-p)^{n-i} - \sum_{i=0}^{r-1} \binom{n}{i} p^i (1-p)^{n-i}. \end{aligned} \tag{1.5.2}$$

在实际工作中, 我们可以选取最大的 r 和最小的 s, 使得

$$\sum_{i=0}^{r-1} \binom{n}{i} p^i (1-p)^{n-i} \leqslant \frac{\alpha}{2}, \tag{1.5.3}$$

$$\sum_{i=0}^{s} \binom{n}{i} p^i (1-p)^{n-i} \geqslant 1 - \frac{\alpha}{2}. \tag{1.5.4}$$

因此

$$P\left\{X_{(r)} \leqslant \xi_p \leqslant X_{(s)}\right\} \geqslant 1 - \frac{\alpha}{2} - \frac{\alpha}{2} = 1 - \alpha.$$

对于单侧置信区间 $[X_{(r)}, \infty)$ 或 $(-\infty, X_{(s)}]$, 选 r 或 s 时, 只需将式 (1.5.3) 和式 (1.5.4) 中的 $\alpha/2$ 换为 α 即可.

当 $n \leqslant 20$ 时, 对于给定的 p 和 α, 查二项分布数值表 (附表 2) 可以得到满足式 (1.5.3) 的最大整数 r 和满足式 (1.5.4) 的最小整数 s. 当 n 相当大而 p 不太接近于 0 或 1 时, 可以用正态分布逼近式 (1.5.3) 左边之和, 即

$$\sum_{i=0}^{r-1} \binom{n}{i} p^i (1-p)^{n-i} \approx \Phi\left(\frac{r - 0.5 - np}{\sqrt{np(1-p)}}\right),$$

此处作了连续性修正. 因此, 由式 (1.5.3) 可得

$$\Phi\left(\frac{r - np - 0.5}{\sqrt{np(1-p)}}\right) \leqslant \frac{\alpha}{2}.$$

由此可取

$$r = \lfloor np + 0.5 + z_{\alpha/2}\sqrt{np(1-p)}\,\rfloor, \tag{1.5.5}$$

其中 $z_{\alpha/2}$ 是标准正态分布的 $\alpha/2$ 分位数, $\lfloor x \rfloor$ 表示小于等于 x 的最大正整数. 同理可取

$$s = \lceil np - 0.5 + z_{1-\alpha/2}\sqrt{np(1-p)}\,\rceil. \tag{1.5.6}$$

其中 $\lceil x \rceil$ 表示大于等于 x 的最小正整数.

经过上述方式定出的置信区间 $[X_{(r)}, X_{(s)}]$, 其置信水平不低于 $1-\alpha$, 但可以大于它, 因此, 一般来说, 这种方法偏于保守. 但它与大样本区间估计相比较, 不涉及密度估计所带来的麻烦, 使用上很方便.

例 1.5.2 从某工厂的产品仓库中随机取 16 个零件, 测得它们的长度 (单位: cm) 为:

2.14, 2.10, 2.13, 2.15, 2.13, 2.12, 2.13, 2.10,
2.15, 2.12, 2.14, 2.10, 2.13, 2.11, 2.14, 2.11.

求该零件长度分布的中位数的置信水平为 0.95 的置信区间.

解 由题意可知, $n=16$, $p=0.5$, $\alpha = 1-0.95 = 0.05$. 查二项分布数值表 (附表 2) 可得

$$\sum_{i=0}^{3} \binom{n}{i} p^i (1-p)^{n-i} = 0.0106 < 0.025,$$

$$\sum_{i=0}^{4} \binom{n}{i} p^i (1-p)^{n-i} = 0.0384 > 0.025,$$

$$\sum_{i=0}^{11} \binom{n}{i} p^i (1-p)^{n-i} = 0.9616 < 0.975,$$

$$\sum_{i=0}^{12} \binom{n}{i} p^i (1-p)^{n-i} = 0.9894 > 0.975.$$

于是, 最大的整数 $r=4$, 最小的整数 $s=12$. 因此可得 $X_{(4)} = 2.11$, $X_{(12)} = 2.14$. 故中位数的置信水平为 0.95 的双侧置信区间为 $[2.11, 2.14]$.

如果用公式 (1.5.5) 和公式 (1.5.6), 则可得相同的结果. 下面我们进行计算. 查标准正态分布数值表可得 $z_{\alpha/2} = z_{0.025} = -1.96$. 因此, 由式 (1.5.5) 可得

$$r \approx 16 \times 0.5 + 0.5 - 1.96 \times \sqrt{16 \times 0.5 \times 0.5} = 4.58,$$

同理由式 (1.5.6) 可得 $s \approx 11.42$. 于是取 $r=4$, $s=12$, 可得 $X_{(4)} = 2.11$, $X_{(12)} = 2.14$. 故中位数的置信水平为 0.95 的双侧置信区间为 $[2.11, 2.14]$.

1.6 习 题 1

1.1 设总体 X 具有分布函数 $F(x)$ 和概率密度函数 $f(x)$, $X_1, \cdots, X_n$ 是来自 X 的独立同分布样本, 其次序统计量为 $X_{(1)}, \cdots, X_{(n)}$, 求极差 $R = X_{(n)} - X_{(1)}$ 的密度函数.