基于排序集样本的非参数检验和估计

JIYU PAIXUJI YANGBEN DE FEICANSHU JIANYAN HE GUJI

张良勇 著

Nonparametric Test and Estimation Based on Ranked Set Samples

河北科学技术出版社

张良勇，河北经贸大学副教授，北京理工大学理学博士。主要研究方向为统计推断、非参数统计、排序集抽样理论等。已在国内外发表学术论文20余篇，其中SCI和EI期刊检索论文近10篇；主持河北省教育厅青年拔尖人才计划项目1项，河北经贸大学科研基金项目1项，并以前三身份参与国家级项目和省级自然科学基金项目6项。

基于排序集样本的非参数检验和估计

JIYU PAIXUJI YANGBEN DE FEICANSHU JIANYAN HE GUJI

张良勇 著

Nonparametric Test and Estimation Based on Ranked Set Samples

河北科学技术出版社

图书在版编目（CIP）数据

基于排序集样本的非参数检验和估计／张良勇著
. --石家庄：河北科学技术出版社，2016.8
ISBN 978-7-5375-8546-0

Ⅰ.①基… Ⅱ.①张… Ⅲ.①非参数检验-研究
Ⅳ.①O212.1

中国版本图书馆 CIP 数据核字(2016)第 190298 号

基于排序集样本的非参数检验和估计

张良勇 著

出版发行 河北科学技术出版社
地　　址 石家庄市友谊北大街 330 号（邮编：050061）
印　　刷 石家庄燕赵创新印刷有限公司
开　　本 787×1092 1/16
印　　张 11
字　　数 260 千字
版　　次 2016 年 8 月第 1 版
2016 年 8 月第 1 次印刷
定　　价 36.00 元

■国家自然科学基金数学天元资助项目(项目编号:11426083)

■河北省自然科学基金资助项目(项目编号:A2015207012)

■河北经贸大学学术著作出版基金资助项目(博士学位论文)

■河北省高等学校科学技术研究青年基金项目(项目编号:QN2015153)

■河北经贸大学科研基金项目(项目编号:2015KYQ03、2016KYQ08)

■河北省重点学科应用统计学资助

前　言

统计学的基本问题是利用观测的样本来推断总体的一些性质,样本的好坏直接影响推断结果的准确性. 因此,在统计学中,如何有效地获得样本是我们所关注的对象,通常我们采用简单随机抽样方法获得统计所需的样本. 然而,当感兴趣的样本测量比较耗时、耗费时,我们就需要寻找更为简便有效的抽样方法.

排序集抽样是一种提高抽样效率的方法,适合于样本实际测量困难但主观排序容易的场合. 由排序集抽样得到的排序集样本不仅包含了样本信息,还包含了次序信息,因此与简单随机样本相比,在同等样本量下排序集样本所含的信息更加丰富,样本更具有代表性. 近几十年,排序集抽样方法被广泛应用到环境科学、医学、农学和经济学等方面. 随着排序集抽样方法在现代生活中的广泛应用,有关排序集抽样的统计推断也成为当今统计学研究的一个热点.

本书取材是根据作者博士论文和项目研究过程中对该领域的研究成果及所积累的资料撰写而成,其中相当一部分内容是最新成果,反映了该领域的现代面貌. 主要研究内容分以下几个方面:

(1) 建立了基于非均等排序集抽样的符号检验统计量,确定了新符号检验统计量的具体分布,证明了新符号检验统计量的渐近正态性,给出新符号检验统计量的 Pitman 效率因子和功效函数. Pitman 渐近相对效率和检验功效的研究结果表明,新符号检验优于基于简单随机抽样和排序集抽样的符号检验.

(2) 建立了基于非均等排序集抽样的加权符号检验统计量,证明了在原假设条件下加权符号检验统计量的分布不依赖于总体分布,证明了加权符号检验统计量的渐近正态性,确定了使加权符号检验的 Pitman 效率因子

达到最大的权重向量,并证明出最优权重向量不依赖总体分布. Pitman 渐近相对效率和模拟功效的研究结果表明,基于非均等排序集抽样的最优加权符号检验优于基于排序集抽样的最优加权符号检验.

(3) 建立了基于非均等排序集抽样的符号秩检验统计量,证明了在原假设条件下新符号秩检验统计量的分布具有对称性且不依赖于总体分布,证明了新符号秩检验统计量的渐近正态性,给出新符号秩检验的 Pitman 效率因子. Pitman 渐近相对效率和模拟功效的研究结果表明,新符号秩检验优于基于简单随机抽样和排序集抽样的符号秩检验.

(4) 建立了基于非均等排序集抽样的秩和检验统计量,证明了在原假设条件下新秩和检验统计量的分布具有对称性且不依赖于总体分布,证明了新秩和检验统计量的渐近正态性,给出新秩和检验的 Pitman 效率因子. Pitman 渐近相对效率和模拟功效的研究结果表明,新秩和检验优于基于简单随机抽样和排序集抽样的秩和检验.

(5) 建立了基于非均衡排序集样本的分位数符号检验,首先考虑均衡排序集样本的分位数符号检验,其次给出非均衡排序集样本的分位数符号检验,然后在不完美非均衡抽样下考虑符号检验,给出了统计量的期望和方差,证明了其适应任意分布和渐近正态性,并讨论了相对于均衡抽样的 Pitman 效率. 接着,考虑了基于不完美非均衡抽样的与次序有关的加权符号检验,分析证明了具有较高 Pitman 效率的适应任意分布的最优权数. 根据最优权数性质和其他一些理论证明,具体给出不同分位数的最优抽样,弥补了排序集抽样在检验分位数上的不足. 随后,研究了排序集样本的最优挑选设计,又进一步给出中位数排序集样本的符号检验和非均衡排序集样本的分布检验等内容.

(6) 建立了基于排序集样本的调查费用分析,显示费用要小于简单随机样本,然后考虑了基于广义排序集抽样的可靠度估计,研究结果表明:当抽样费用和排序费用不能忽略不计时,每个排序小组中抽测两个或更多样本单元的抽样方案具有较高的费用效率.

(7) 针对生存函数的估计问题,建立了排序集样本下随机删失数据的

乘积限估计量,证明了估计量的强相合性,确定了其强收敛速度,模拟计算和实际应用的结果均表明排序集抽样效率高于简单随机抽样.

(8) 建立了基于随机删失排序集样本的核密度估计量,证明了新估计量的渐近正态性,并通过计算机模拟来比较排序集抽样估计量与相应的简单随机抽样下估计量的相对效率,模拟结果表明排序集抽样优于简单随机抽样.

本书的研究具有重要的理论和实用价值,研究结果为排序集样本下其他的统计推断如经验似然、区间估计、位置参数 Hodges – Lehmann 估计及回归估计等方面提供了研究的基础和思路. 另外,研究结果在临床医学、环境科学、经济学等领域都有广泛的应用前景.

本书的出版得到了国家自然科学基金数学天元资助项目(项目编号:11426083)、河北省自然科学基金资助项目(项目编号:A2015207012)、河北经贸大学学术著作出版基金资助项目(博士学位论文)、河北省高等学校科学技术研究青年基金项目(项目编号:QN2015153)、河北经贸大学科研基金项目(项目编号:2015KYQ03、2016KYQ08)及河北省重点学科应用统计学的资助,同时也得到我的妻子董晓芳的帮助,作者谨在此一并表示感谢.

张良勇

2016 年 7 月

内容简介

排序集抽样是一种提高抽样效率的方法,适合于样本实际测量困难,但主观易于排序的场合. 由排序集抽样得到的排序集样本不仅包含了样本信息,还包含了次序信息,因此与简单随机样本相比,在同等样本量下排序集样本所含的信息更加丰富,样本更具有代表性. 近几十年,排序集抽样方法被广泛应用到环境科学、医学、农学和经济学等方面. 随着排序集抽样方法在现代生活中的广泛应用,有关排序集抽样的统计推断也成为当今统计学研究的一个热点.

本书主要研究的内容有:基于非均等排序集抽样的非参数检验,包括符号检验、加权符号检验、符号秩检验和秩和检验,重点给出检验统计量的建立,统计量性质的分析和检验的功效;基于非均衡排序集样本的分位数符号检验,包括均衡与非均衡排序集样本的分位数符号检验,基于排序集样本的最优挑选设计,中位数排序集抽样的符号检验和基于非均衡排序集样本的分布检验;基于排序集抽样的费用分析,包括基于排序集抽样方法的调查费用分析和费用模型下排序集样本的可靠度估计;基于删失排序集样本的乘积限估计和核密度估计,分别包括乘积限估计量和核密度估计量的构建、估计量性质和估计模拟效率的分析.

目　　录

第一章　绪论

在统计中有效的抽样方法是我们所关注的对象,尤其是当感兴趣的样本测量比较耗时、耗费时,我们就需要寻找更为简便有效的抽样方法. 统计学的基本问题是利用观测的样本来推断总体的一些性质,样本的好坏直接影响推断结果的准确性. 因此,在统计学中,如何有效地获得样本是我们所关注的对象,通常我们采用简单随机抽样方法获得统计所需的样本. 然而,当感兴趣的样本测量比较耗时、耗费时,我们就需要寻找更为简便有效的抽样方法.

排序集抽样是一种提高抽样效率的方法,适合于样本实际测量困难,但主观排序容易的场合. 由排序集抽样得到的排序集样本不仅包含了样本信息,还包含了次序信息。因此与简单随机样本相比,在同等样本量下排序集样本所含的信息更加丰富,样本更具有代表性. 近几十年,排序集抽样方法被广泛应用到环境科学、医学、农学和经济学等方面. 随着排序集抽样方法在现代生活中的广泛应用,有关排序集抽样的统计推断也成为当今统计学研究的一个热点话题.

本书主要研究了基于排序集样本的非参数检验和估计问题. §1.1 首先概述了非参数统计方法的主要特点,其次简述次序统计量及其分布、分位数定义,然后介绍几种常见非参数估计方法,非参数检验的概念与特点、非参数检验的评价标准. §1.2 简述了基于简单随机抽样的非参数检验,包括符号检验、符号秩检验、秩和检验. §1.3 阐述了排序集抽样方法,并介绍该抽样下非参数检验的国内外研究成果. §1.4 阐述了非均等排序集抽样方法. §1.5 阐述了非均衡排序集抽样方法. §1.6 阐述了随机删失排序集抽样方法.

§1.1 非参数统计

非参数统计是相对于参数统计而存在的. 我们知道,数理统计中的参数统计研究了这样的问题:假定总体的理论分布类型已知,而分布中的若干参数未知,于是我们通过抽出的样本来估计这些未知参数,并用假设检验方法来检验这些未知参数的合理性. 然而在非参数统计中,总体的理论分布是未知的,是一个有待检验的假设,而实际应用中这种问题又是非常普遍的,即面对一大堆试验所得的数据,它们服从何种分布往往是不知道的. 非参数统计方法就是致力于解决理论分布未知时的检验与估计问题.

§1.1.1 非参数统计方法的主要特点

一般来说,处理非参数统计问题的方法,我们称之为非参数统计方法. 非参数统计方法与参数统计方法的优劣性大体上是互补的,即一者的优点往往就是另一者的弱点. 与参数统计方法相比,非参数统计方法有如下几个特点:

(1) 非参数统计方法的适用面广. 非参数统计方法无需假定总体是在何种类型分布条件下,即可仅通过样本所获得的信息对所关心的问题做出判断. 而参数统计方法必须预先假定总体具有某种分布类型,对一群纯数据来讲,这个要求通常是过高的,且如果假定理论分布类型与实际分布不符,则做出的判断会与实际差之甚远. 因而非参数统计方法比参数统计方法适用面宽.

(2) 样本方法是非参数统计的基本方法. 统计量的分布总是依赖于总体分布,即建立估计与检验统计量时,通常是将它的极限分布作为总体分布. 有时其精确分布过于复杂时,我们宁愿用其极限分布代替精确分布.

(3) 非参数统计方法只采用了样本的“一般”信息,而扔掉了其“特殊”信息. 一般来说,非参数统计方法主要是根据实际数据中蕴含的“一般”信息来构成估计量或检验统计量,用于估计与假设检验,而缺乏对特殊分布情况的针对性. 例如,样本均值 $\overline{X}$ 肯定提供了总体期望 μ 的信息,所以可以用作 μ

的估计. 对如此总体为正态分布或接近正态分布时,则其抽取的数据多为“两头低,中间高”的情况,如此平均值 $\overline{X}$ 较为可信. 但如果总体分布形式并不规则时,用数据的平均值反映总体分布的期望 μ 就不合理了.

(4) 非参数统计具有良好的稳健性. 统计方法的稳健性,是这样的一种性质. 当真实统计模型与假定模型相差不多时,这种统计方法依然具有良好的适用性质. 若真实模型与假定模型稍有差异,这种统计方法就不再适用时,我们就认为此种统计方法不具备稳健性质. 非参数统计方法对总体模型的限制少,一般并不假定其具有何种形状,仅依据样本中的“一般”信息对总体的形状、特征进行估计. 可见,当总体模型稍有变动时,对这种估计并无太大影响.

总的来说,非参数统计方法具备一些参数统计方法无法比拟的优点,同时也暴露出一些弱点. 故而在实际应用中,不应拘于单纯使用参数或非参数统计方法,只应参照尽可能多的能收集到的信息,采用最为接近实际方法,以达到预期的估计、检验、预测与控制等目的.

§1.1.2 次序统计量及其分布

次序统计量是统计学中最常见的统计量,在非参数统计中占有极其重要的地位,故此简述一下次序统计量的概念.

定义 1.1　假定总体 X 有容量为 n 的样本 $X_1, X_2, \cdots, X_n$,将 $X_1, X_2, \cdots, X_n$ 按从小到大排序后生成的统计量

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

则称统计量 $\{X_{(1)}, X_{(2)}, \cdots, X_{(n)}\}$ 为次序统计量. 其中 $X_{(i)}$ 是第 i 个次序统计量. 次序统计量是非参数统计的理论基础之一,许多非参数统计量的性质与次序统计量有关.

如果总体分布函数为 $F(x)$,则次序统计量 $X_{(r)}$ 的分布函数为

$$F_{(r)}(x) = P(X_{(r)} \leqslant x) = P(\text{至少 } r \text{ 个 } X_i \text{ 小于或等于 } x)$$

$$= \sum_{i=r}^{n} \binom{n}{r} F^i(x)[1 - F(x)]^{n-i}.$$

如果总体分布密度为 $f(x)$,则次序统计量 $X_{(r)}$ 的密度函数为

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

次序统计量 $X_{(r)}$ 和 $X_{(s)}$ 的联合密度函数为

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

§1.1.3 分位数的定义

一组数据从小到大排序后,每一个数在数据中的序非常重要,给定序,寻找对应的数据,用分布的语言来说,就是找分位数.

不失一般性,对任意分布而言,分布的分位数定义如下.

定义 1.2　假定 X 服从概率密度为 $f(x)$ 的分布,令 $0<p<1$,满足等式 $F(m_p)=P(X<m_p)\leqslant p$,$F(m_p+)=P(X\leqslant m_p)\geqslant p$ 唯一的根 m_p 称为分布 $F(x)$ 的 p 分位数.

定义 1.3　假定 X 服从概率密度为 $f(x)$ 的分布,令 $0<p<1$,满足等式 $F(x)=P(X<m_p)=p$ 的唯一的 m_p 称为分布 $F(x)$ 的 p 分位数.

分位数是刻画分布的重要特征,经验分布函数的基本思想就是建立在分位数估计上的. 如果一组数据有 n 个值,分布的第 $i/(n+1)$ 分位点的估计由第 i 小数据生成. 一般而言,对任意分位数可以构造如下估计.

给定 n 个值 $X_1,X_2,\cdots,X_n$,可以根据下面的公式计算任意 p 分位数的值:

$$m_p=\begin{cases}X_{(k)}, & \dfrac{k}{n+1}=p,\\ X_{(k)}+(X_{(k+1)}-X_{(k)})[(n+1)p-k], & \dfrac{k}{n+1}<p<\dfrac{k+1}{n+1}.\end{cases}$$

§1.1.4 几种常见非参数点估计方法

我们假定简单样本 $X_1,X_2,\cdots,X_n$ 来自总体 X,X 具有连续的分布函数 $F(x)$,而 $F(x)$ 为未知,此时,利用样本 $X_1,X_2,\cdots,X_n$ 去估计总体分布 $F(x)$ 的某些重要参数,如 p 分位数、对称中心、位置差、刻度差,就显得极为重要. 如在参数统计中的定义一样,我们将不带任何未知参数的统计量 $g(X_1,X_2,\cdots,$

X_n)，观测值 $g(x_1,x_2,\cdots,x_n)$ 作为某参数 θ 的估计. 我们称 $\hat{\theta}=g(X_1,X_2,\cdots,X_n)$ 为 θ 的点估计量，而 $\hat{\theta}=g(x_1,x_2,\cdots,x_n)$ 为 θ 的点估计值，统称为 θ 的点估计.

1. 总体分位数的点估计

设 ξ_p 为分布函数 $F(x)$ 的唯一 p 分位数，则有 $F(\xi_p)=p$. 可以证明，当 $F(x)$ 为严格单调连续函数时，其 p 分位 ξ_p 是唯一存在的. 对于 ξ_p 的估计，我们自然想到利用样本 p 分位数来估计，即取 ξ_p 估计为

$$\hat{\xi}_p=m_{np}=\begin{cases}X_{([np]+1)}, & np\text{ 为非整数}\\ \dfrac{1}{2}[X_{(np)}+X_{(np+1)}], & np\text{ 为整数}\end{cases}\tag{1.1}$$

特别地，当 $p=0.5$ 时，总体中位数 $\xi_{0.5}$（$F(\xi_{0.5})=0.5$）的点估计记为

$$\hat{\xi}_{0.5}=m_n=\begin{cases}X_{(k+1)}, & n=2k+1,\\ \dfrac{1}{2}[X_{(k)}+X_{(k+1)}], & n=2k.\end{cases}$$

可以证明关于 $\hat{\xi}_p$ 有如下结果：

定理 1.1　设简单样本 $X_1,X_2,\cdots,X_n$ 来自总体 X，ξ_p 为连续型总体分布函数的 p 分位数，m_{np} 为(1.1)式定义的样本 p 分位数，则有

$$P\{\lim_{n\to\infty}m_{np}=\xi_p\}=1.$$

2. 对称中心的点估计

我们称 X 的分布函数 $F(x)$ 是关于原点对称的，即指 $F(x)$ 满足

$$F(-x)=1-F(x)$$

或密度函数 $f(x)$ 满足

$$f(-x)=f(x)$$

可见，当 $F(x)$ 关于原点对称，则有 $F(0)=\dfrac{1}{2}$.

一般地，设 $F(x)$ 为关于原点对称的分布函数，则记 $F(x-\theta)$ 为关于参数 θ 对称的分布函数，称 θ 为 F 的对称中心.

当总体分布为 $F(x-\theta)$，函数 $F(x)$ 为关于原点对称时，常用的估计统