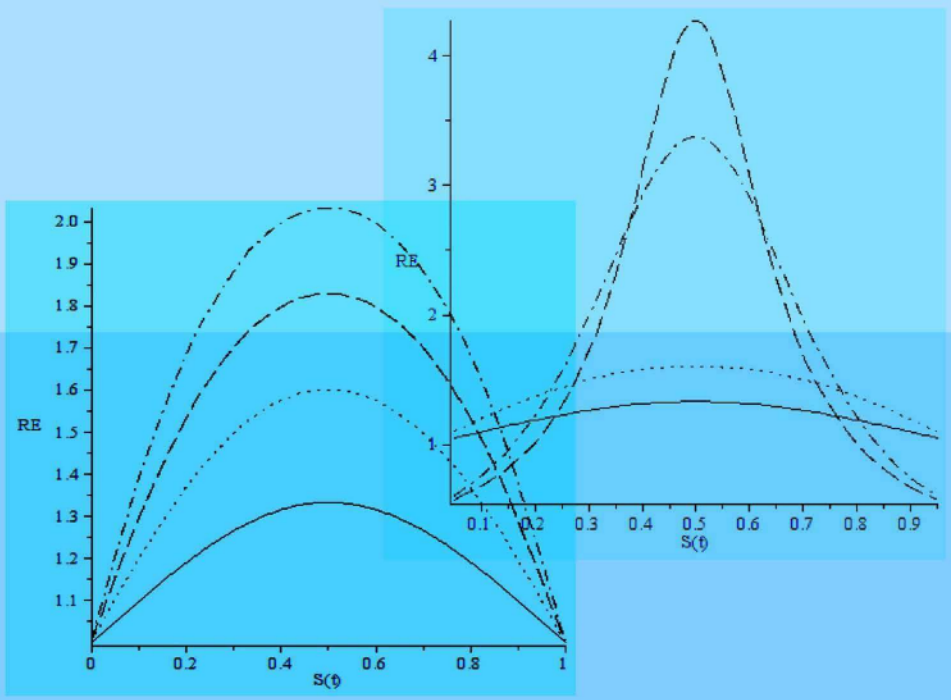
排序集抽样下
生存数据的非参数估计

PAIXUJI CHOUYANGXIA
SHENGCHUN SUJU DE FEICANSU GUJI

董晓芳 著

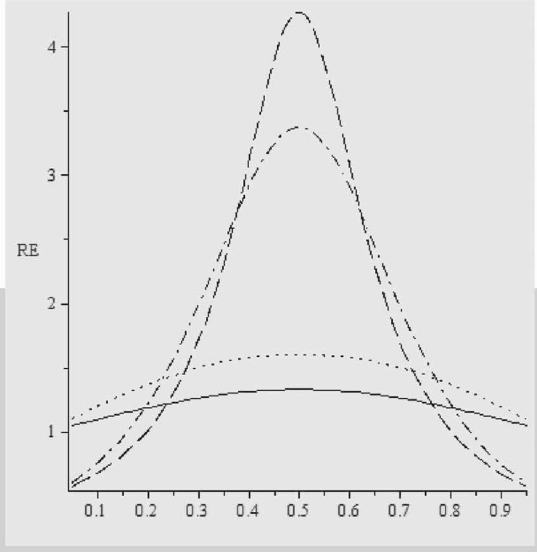


河北科学技术出版社

排序集抽样下生存数据的非参数估计

PAIXUJI CHOUYANGXIA SHENGCUN SHUJU DE FEICANSHU GUJI

董晓芳 著



河北科学技术出版社

图书在版编目（CIP）数据

排序集抽样下生存数据的非参数估计 / 董晓芳著
. -- 石家庄 : 河北科学技术出版社, 2016.8
ISBN 978-7-5375-8548-4

Ⅰ. ①排… Ⅱ. ①董… Ⅲ. ①序贯分析-非参数法-研究 Ⅳ. ①O212.3

中国版本图书馆 CIP 数据核字(2016)第 189335 号

排序集抽样下生存数据的非参数估计

董晓芳 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

前　言

生存数据不仅出现在生物医学中，而且出现在工业可靠性、社会科学和商业研究中，因而生存数据的统计分析有着广泛的应用价值，日益受到人们的关注．经典统计采用简单随机抽样获得所需的生存数据．然而，由于受到实验规模和试验费用的限制，对大量随机样本的实际测量可能比较困难．排序集抽样是一种提高抽样效率的方法，适合于样本易于排序但不易于数量化的场合，此抽样方法已被广泛应用到农业、医学、生态环境和经济等领域．为保证抽取的生存数据既有一定的代表性，又尽可能减少抽样检查费用，本书针对排序集抽样下生存函数、平均寿命和分位寿命的估计问题进行了研究．本书的研究工作主要有以下几个方面：

第一，针对完全生存数据下总体生存函数的估计问题，建立了均衡排序集样本经验生存函数，证明了估计量的无偏性和渐近正态性，并通过方差的比较，证明出它一致优于简单随机样本经验生存函数；在两个典型的非均衡排序集抽样下，利用样本经验生存函数去估计生存函数，证明了估计量的渐近正态性，并给出了两个抽样方案适用的范围．

第二，针对删失生存数据下总体生存函数的估计问题，建立了生存函数的直接乘积限估计和平均秩乘积限估计，证明了直接估计的自相容性和平均秩估计的强相合性，模拟结果表明了平均秩估计优于直接估计，而它们都优于删失简单随机样本的乘积限估计．

第三，针对完全生存数据下总体平均寿命的估计问题，建立了基于非均衡排序集样本的加权无偏估计，证明了估计量的渐近正态性，给出极端抽样和中位数抽样下估计的权数，并通过方差的比较，证明了这两种抽样效率均高于简单随机抽样.

第四，针对删失生存数据下总体平均寿命的估计问题，利用生存函数的乘积限估计构造了平均寿命的估计量，给出了估计差值的渐近表示，模拟结果表明了删失排序集抽样效率高于删失简单随机抽样.

第五，针对完全生存数据下总体分位寿命的估计问题，建立了均衡排序集样本分位寿命，证明了估计量的强相合性和渐近正态性，并通过渐近方差的比较，证明了它一致优于简单随机样本分位寿命；在非均衡排序集抽样下，建立了与秩次有关的加权估计，证明了估计量的强相合性和渐近正态性，给出了在固定抽样方案下使渐近方差达到最小的权数，再利用最优权数的适应任意分布性，确定出使估计效率达到最大的抽样方案，渐近相对效率的计算结果表明了最优抽样的高效率性；针对总体中位寿命的估计问题，分析了中位排序集样本中位寿命的统计性质，渐近相对效率的比较结果表明：中位数排序集抽样方案一致优于简单随机抽样和均衡排序集抽样.

第六，针对删失生存数据下总体分位寿命的估计问题，建立了删失排序集样本的分位寿命，模拟结果表明了删失排序集样本分位寿命的估计效率优于删失简单随机样本分位寿命.

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. 本书的研究有重要的理论和实践意义，所得结果在管理科学、临床医学和可靠性工程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：2016KYQ08，2015KYQ03）及河北省重点学科应用统计学的资助，同时也感谢我的丈夫张良勇的帮助，作者谨在此一并表示感谢.

内容简介

生存分析是对生存数据进行统计分析的一门学科，一直受到国内外统计学家的关注，研究异常活跃．在经典统计中我们通常采用简单随机抽样的方法获得所需的生存数据．然而，由于受到实验规模和试验费用的限制，对大量随机样本的实际测量可能比较困难．但为了保证抽样样品的质量指标既有一定的代表性，又尽可能减少抽样检查费用，我们需要寻找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，适合于样本易于排序但不易于数量化的场合，随着排序集抽样方法在医学、农业、经济和生态环境等领域的广泛应用，基于排序集抽样的生存分析，成为近年来国内外专家和学者研究的热点问题之一，并且在此领域的理论和应用上都有了一定的进展．

本书主要研究的内容有：基于均衡排序集抽样下的生存数据的非参数估计，包括生存函数估计、平均寿命估计、分位寿命估计、对称分布的修正估计；基于非均衡排序集抽样下生存数据的非参数估计，包括生存函数估计、平均寿命估计，分位寿命估计、中位寿命估计；基于删失排序集抽样下的生存数据的非参数估计，包括生存函数直接乘积限估计和平均秩乘积限估计、平均寿命估计、分位寿命估计；完全数据和删失数据的实例分析．

■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:2016KYQ08,2015KYQ03)

■河北省重点学科应用统计学的资助

目　录

第1章 绪 论

1.1 研究的目的和意义

生存分析是对生存数据进行统计分析的一门学科，它是根据医学、生命科学、可靠性工程、保险等科学研究中的大量问题提出的[1-3]. 近三十年来生存分析受到国内外统计学家的关注，研究异常活跃. 生存数据不仅出现在生物医学中，而且出现在可靠性工程、犯罪学、社会学、市场学和商业研究中. 在这些领域生存数据的例子有：可靠性工程电子设备（元件或系统）的寿命，犯罪学中重罪犯人的假释时间，社会学中首次婚姻的持续时间. 它也可以不是时间，它可以是汽车车轮转动的圈数，也可以是市场学中报纸或杂志的篇幅和订费，甚至可能是工人们的补偿索赔等. 生存分析含有许多实用的方法和丰富的理论. 随着医疗实践和工程实践及其他领域的推动，不断有新的统计方法出现，应用范围越来越广. 在经典统计中我们通常采用简单随机抽样（Simple Random Simpling，简称为 SRS）的方法获得所需的生存数据. 然而，获得牢靠的生存数据不是一件容易的事. 一方面，由于受到实验规模和试验费用的限制，我们不可能对大量的随机样本进行实际测量；另一方面，对样本的实际测量可能比较困难或者具有破坏性. 为保证抽样样品的质量指标既有一定的代表性，又尽可能减少抽样检查费用，我们需要寻找简便有效的抽样方法.

在 20 世纪 50 年代早期，澳大利亚农业学家 Mclntyre[4] 在估计农场上牧草产量时提出了排序集抽样（Ranked Set Sampling，简称为 RSS）的方法. 测量草地的产草量，需要把草割下来，晒干再去称量干草的重量，是

非常消耗时间和劳动的过程，但是有经验的眼睛可以对一组数目较少的几块草地进行相当精确的排序，而不需要进行准确的测量．Mclntyre 采取了如下的抽样机制，首先随机抽取一组大小为 m 的草地，通过肉眼对这组草地的产草量进行由小到大的排序，次序最小的草地被抽出．然后从农场中随机抽取另一组大小为 m 的草地，通过肉眼对他们的产草量进行由小到大的排序，次序为 2 的草地被抽出．依此类推，直至从第 m 组草地中抽出次序最大的草地．最后只是对抽出的 m 个草地进行割草和称重．以上整个过程称为一次循环，这一循环重复 k 次，则得到样本量为 $n=mk$ 的排序集样本．1966 年，Hall & Dell[5] 设计了一个实验，通过排序集抽样方法对阿巴拉契亚山脉橡树林的产量进行了估计，从这之后，排序集抽样方法逐渐被人们重视．在实际中，只要感兴趣的变量不易测量，但较容易用主观经验判断或其他不需要具体测量的方法对样本进行排序时，使用排序集抽样比简单随机抽样更加有效．例如，研究危险废弃物场所的污染程度时，需要测量有毒化学品的污染指标，通常费用会非常昂贵．通过目视观察落叶或土壤的变色，给出变量的排序，再从排序的变量中有选择性抽取一部分样本进行测量，这样可减少抽样次数．1997 年，Yu & Lam[6] 验证了当估计美国内华达测试基地相邻地区表层土壤中钚的含量时，排序集抽样效率比简单随机抽样高．

除了在农业和生态环境上的应用，排序集抽样方法在医学领域也有广泛的应用．人类的许多定量特性，如高血压和肥胖等，遗传度相当高，但遗传机制尚不清楚．这就需要对配对亲属的等位基因测验，并进行遗传相关性分析，通常需要花费大量的金钱和时间来进行实验室检测．然而，医生可以使用排序集抽样技术对病人进行合理地选择，比如依据诸如年龄、体重、身高、血压和健康史等信息对病人进行选择，这一过程的花费是可以忽略的．1995 年，Risch & Zhang[7] 在《Science》上论证了对配对亲属进行极值排序集抽样，遗传相关性试验的效率能得到显著地提高．再例如用双能 X 线吸收法测量人体骨密度水平是花费较高的，但有经验的医生可以不需要做实际测量，对几个被检查的人骨密度水平排序．2004

年，美国俄亥俄州立大学骨密度研究中心[8]采用了 RSS 方法，使得能以较 SRS 方法少的测量个体，来做出对人群的骨密度水平的合理估计．另外，Chen（2007）[9]和 Bouza（2009）[10]分别验证了排序集抽样方法对肺癌和艾滋病临床研究的高效性．

随着排序集抽样方法在医学、农业、经济和生态环境等领域的广泛应用[4~24]，基于排序集抽样的生存分析，成为近年来国内外专家和学者研究的热点问题之一，并且在此领域的理论和应用上都有了一定的进展．然而，大多数文献都是针对生存分析中参数统计进行研究，适当的模型或分布可用来拟合数据或者可以假定数据来自某种分布的总体时，用参数统计方法比较简便，但在实际中有许多情况使得没有一个现实的基础来选择一个特定的分布类型．例如，对一个新试制的治癌药品，事先可能就没有足够的信息来判断病人服用后缓解时间属于哪个分布类．这时我们就无法使用参数统计方法，而是借助非参数方法进行统计推断．非参数统计作为数理统计学的一个分支，是专门研究与分析在总体分布未知的情况下，有关数据总体的统计信息的预测与推断的理论与方法．在过去的 20 年中，随着医学研究中的临床试验快速研究，使生存统计分析方法研究的重点从参数模型转移到非参数模型．此外，以往文献研究的排序集抽样下生存数据都是完全数据，对于试验产生的删失数据的研究还属于空白．但是生命科学中的生存数据有一个最重要的特点：在研究期间结束时某些个体身上还没有出现我们关心的事件．例如，在研究周期结束时某些患者仍然活着或处于缓解状态，这些个体的确切生存时间不知道．在可靠性工程许多研究中，由于种种条件的限制也不可能获得完全样本．例如，受试验时间、费用等的限制，不可能将寿命试验做到所有元件都失效．

生存函数、平均寿命和百分位寿命均是生存分析主要的数量指标，当生存时间的分布类型未知，为使获得的生存数据包含更多生存时间的信息，本书研究排序集抽样下生存数据的非参数估计问题，在均衡排序集抽样、非均衡排序集抽样和随机删失排序集抽样下分别建立生存函数、平均寿命和百分位寿命的非参数估计量，推导相应估计量的性质，比较与简单

随机抽样的相对效率，并把理论结果应用到管理科学和临床医学的实例分析中．目前，国内外对这些问题的研究还属于空白，研究结果不仅能应用到管理科学、医药卫生、可靠性工程，而且在保险数学、犯罪学、社会学、市场学、环境科学等高科技领域都有广泛的应用前景．

1.2 国内外研究现状及发展趋势

1.2.1 均衡排序集抽样下统计分析研究现状

为了便于和后来提出的非均衡抽样方法的区分，Mclntyre[3]提出的排序集抽样方法又称为均衡排序集样本（Balanced Ranked Set Sampling，简称为 BRSS），因为此样本包含每一个次序统计量的信息相同．对均衡排序集样本进行统计推断的理论上的第一个结果是 Takahasi 等[25]得出的，他们证明排序集样本均值是总体均值的无偏估计，并且此估计量的方差要小于简单随机样本的方差．Stockes[26]考虑了用排序集样本来估计总体的方差，结果表明，排序集抽样的精度比简单抽样高．早期文献主要集中在非参数统计分析上，另外还有文献[27～30]．

20 世纪 90 年代初是排序集抽样理论和应用发展的一个转折点，从那以后，关于排序集抽样的各种参数和非参数统计分析被研究，许多有关排序集抽样的新定义和新定理被提出和证明．在基于排序集抽样的参数统计方面，文献[31 ～ 37]研究了参数的极大似然估计和贝叶斯估计；文献[38～41]对参数的假设检验进行了研究；文献[42～47]分析了线性回归模型中的参数估计．在基于排序集抽样的非参数统计方面，文献[48 ～ 50]考虑了总体均值和方差的非参数估计；文献[51 ～ 54]分别首次研究了未知总体中位数的 Mann－Whitney－Wilcoxon 检验、符号检验、符号秩检验和 U 检验；文献[55]首次考虑总体比率的非参数估计；文献[56 ～ 57]给出了未知总体密度函数和分布函数的非参数估计．这些文献结果均表示排序集抽样效率高于简单随机抽样．

进入21世纪，基于均衡排序集抽样的生存分析，开始受到各国学者的重视，并且在此领域进行了大量的研究，在理论和应用上都有很大的发展．关于此领域的研究也成为国内外近年来热点问题之一．

1. 均衡排序集抽样下参数统计研究现状

在参数统计研究中，总体分布类型是已知的，这可以由以往数据分析所积累的经验来判断分布类型．常见的参数分布有正态分布、指数分布、韦布尔分布和对数正态分布等，其中指数分布是生存分析中最重要的一种分布，几乎是专门用于描述电子设备可靠性的一种分布．它计算简单，参数的估计容易，且具有“无记忆性”．有关排序集抽样下参数统计研究主要围绕正态分布和指数分布进行．

生存函数 $S(t)$ 在生存分析中起着重要作用，它表示个体生存时间长于 t 的概率，在可靠性工程上常称 $S(t)$ 为产品的可靠度．2000年，El－Neweihi & Sinha[58]首次考虑了基于BRSS的指数单元可靠度的估计问题．他们指出了BRSS和可靠度理论的关系，样本量为 n 的BRSS中的第 i 个测量值可看作为 n 个相互独立同分布单元组成的表决系统 $i/n(F)$ 的寿命时间．利用这一关系构造了可靠度的一类无偏估计量，证明了这类BRSS估计量的方差都小于简单样本估计量的方差，并从这类无偏估计量中给出方差最小的RSS估计量．2005年，Ghitang[59]进一步肯定了对于指数单元可靠度的估计，排序集抽样一致优于简单随机抽样，但同时通过举例指出文献［58］中的最优估计的方差并不是BRSS中最小的．2006年，Sinha[60]等考虑了服从单参数指数分布的单元可靠度估计问题．此文献首次提出了由BRSS和SRS的次序统计量构造可靠度的无偏估计．通过方差的比较，证明出当样本小组数为2时，BRSS的功效优于SRS，同时也证明了BRSS次序统计量构造的无偏估计比SRS无偏估计功效高．

在估计总体均值方面．2000年，Bhoj[61]讨论了基于BRSS的单参数分布族的均值估计问题，证明了总体生存时间服从指数、Rayleigh或Logistic分布时，用排序集样本均值来估计总体均值的精度要高于简单随机样本均值．2007年，Al－Salen & Al－Ananbeh[62]讨论了基于BRSS的正态分布位