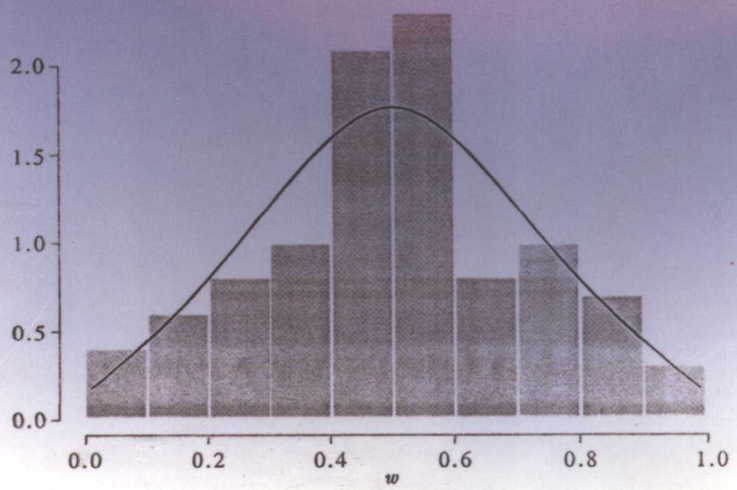
史道济 著

实用极值统计方法

天津大学“211工程”丛书

天津科学技术出版社



实用极值统计方法

史道济 著

天津科学技术出版社

图书在版编目(CIP)数据

实用极值统计方法/史道济著 .—天津:天津科学技术出版社,2006
(天津大学"211 工程"丛书)
ISBN 7-5308-4026-6

Ⅰ.实... Ⅱ.史... Ⅲ.极值(数学)-数理统计-统计方法 Ⅳ.0212

中国版本图书馆 CIP 数据核字(2005)第 108754 号

责任编辑:赵雪慧
版式设计:邱 芳
责任印制:王 莹

天津科学技术出版社出版
出版人:胡振泰
天津市西康路 35 号 邮编 300051 电话(022)23332393(发行部) 23332390(市场部) 27217980(邮购部)
网址:www.tjkjcbs.com.cn
新华书店经销
天津市津通印刷有限公司印刷

开本 787×1092 1/16 印张 13.75 字数 326 000
2006 年 4 月第 1 版第 1 次印刷
定价:25.00 元

前　言

极值统计是专门研究很少发生，然而一旦发生却有巨大影响的随机变量极端变异性的建模及统计分析方法.目前，极值统计的应用已经深入到许多领域.除了传统的水文、气象、地震与工程，近年来在保险、金融以及网络通讯中也得到广泛的应用.本书介绍一些实用的极值统计分析方法，是对国内外有关著作比较全面的总结，希望尽可能地介绍到目前为止已经有的成果，并取名为《实用极值统计方法》.这有二层含义：第一，强调本书重应用，我们并不过多介绍有关极值的概率性质，一系列定理及其证明.当必须给出定理时，一般也不证明，只给出某些相关的参考文献，读者自己可以找到证明；第二，强调统计分析方法，极值统计分析方法区别于一般统计方法，主要在于数据的收集，而不是数据的分析.因此常用的统计分析方法，都仍然适用于极值统计.当然，还有一些是只限于极端数据的分析方法.我们只罗列各种不同方法，没有从理论上比较它们的优良性.

不能奢望通过本书能解决读者所遇到的一切有关极值统计分析问题.一方面，极值理论正处于迅速发展中，有些问题的处理在理论上还没有一种公认为最好的统计方法，有些方法可能还比较粗糙，有待改进，特别对多元极值问题；另一方面，有时认为极值的统计分析不仅是科学方法，而且还是一种“艺术”.对一个实际问题，既要尽可能多地利用包含在数据中的极值信息，又要保持模型的正确性，不能将所有数据都认为是极值，需要在这二者之间进行适当的平衡，这就是一种艺术.

统计分析方法离不开计算机对数据进行各种处理，极值统计也不例外.书中所有的计算程序都是用 R 语言编写的，许多例子会给出各种统计图表，使读者能比较直观地理解.我们之所以选择 R 语言，而不是许多读者所熟识的 Fortran、C、Matlab、S-plus 或 SAS、SPSS 等语言，原因是 R 语言是在 S，S-plus 的基础上发展起来的，属于 GNU 系统的一个自由、免费、源代码开放的软件，是专用于统计计算，具有强大作图功能的优秀工具，且得到了全世界许多优秀的统计计算专家的支持，比商业化软件更能紧跟最新发展的形势.相关的 R 程序及许多数据库都可从附录提供的互联网址上免费下载.当然，也有用其他语言编写的程序，例如 evim、evis、evir 就是分别用 Matlab、S-plus、R 写的极值统计软件包，使用都非常方便.多元极值理论是建立在一元极值理论的基础上，对其进行比较广泛的研究，使其真正得到重视和发展，不过短短十几年的时间.因此，专门的统计分析方法并不很多，许多统计方法还处在探讨之中，不很成熟.本书介绍的方法有些是我们自己的成果，供大家参考，欢迎读者提出中肯的意见.

本书不是为专业统计工作者写的，它适用于各个领域的技术人员及管理人员.现在国内工学、理学、生物学、药学、管理学、经济学等专业的大学本科毕业生大都具有基本概率统计知识.如果对极值统计感兴趣，就可以读懂本书.

书后给出了广泛的参考文献，为对极值理论有兴趣的读者提供进一步阅读的材料.由于本书篇幅有限，不可能非常完善、详尽地给出每一种统计分析方法，有些细节读者可以从所列参考文献中找到.

梁冯珍、邸男、李胜朋、王爱莉、钟君、尹剑、罗俊鹏、关静、谢中华、张春英、沈菲参加了本书部分章节的编写及有关研究.这里要特别感谢 R. L. Smith、J. A. Tawn、S. G. Coles，不仅因为本书介绍的许多实例中的数据是由他们提供的，更因为本人在访英期间得到过他们的许多帮助.感谢 A. Stephenson，我们使用了由他提供的专用于极值统计分析的 R 程序.

作者学识有限，对书中出现的错误或不妥，敬请读者和同行批评指正.

史道济于天津大学理学院

2005 年 3 月

目　录

1 序 言

本书主要研究很少发生，然而一旦发生却产生极大影响的随机事件. 例如洪水、干旱、地震、飓风等自然灾害，经济、金融领域内某些现象的重大变化，导致某个系统失效的随机冲击等. 极端事件有时比正常情况更重要. 历史上，在人们与自然灾害的长期斗争中，保存了不少有关记载，这些珍贵资料记录了曾经发生过的极端事件，而更多的正常现象未必被记录或保存下来. 如何从这些资料预测某种等级自然灾害发生的可能性，是极值理论所研究的问题之一. 在设计水利水文工程时，设计者最关心洪水及干旱的发生. 金融市场的风险是由利率、汇率、股市或商品价格的波动引起的，文献[232]提供了许多由于上述因子急剧变化导致国际大公司破产的例子. 检测一个城市的大气环境质量，需要几个监测点同时监测各种指标，如二氧化硫、氮氧化物、总悬浮颗粒物、一氧化碳等是否在某个水平(如大气环境质量国家标准)以下. 高层建筑要经受来自各个方向的大风，这是设计师必须要考虑的因素. 类似地，任何一个机械零件的设计也必须考虑各种不同类型的强度要求；长江三峡大坝或防汛工程的设计高度必须超过某个可能的最高水位；打破各项体育运动成绩的记录也是一件不寻常的事情；如何设计再保险产品等，都是与极值有关的问题. 可见极值统计在水文、气象、地震、工程、环境、体育及金融、保险、管理等方面有重要应用.

预测是许多科学工作者的一项重要工作. 预测就是以对过去的探讨来求得对未来的了解. 如果要推断某种从未发生过的极端事件在未来发生的可能性，在已掌握的资料中没有记录到如此严重的极端现象，就可由极值理论提供的可以外推的模型来解决这类问题. 尽管对模型的外推，一般持谨慎态度，甚至持批评态度，但迄今为止，极值模型应该是最好方法，还未曾受到严重的挑战.

1.1 什么是极值

在研究极值之前，首先要搞清一个最根本的问题：什么是极值，或者什么样的极值是我们感兴趣的. 从概率意义上讲，极值表示随机变量的极端变异性；从统计意义上讲，极值是指数据集合中的最大值或最小值. 因此每个数据集都有极值，尽管极值与集合中其他数据的差别可能不大. 实际上，这个差别取决于数据集合的规模——样本量，数据集越大，其中的数据越多，最大值就越大，而最小值就越小. 但在严格意义上，什么是极值是一个很难说清楚的问题，就如什么是异常值一样. V. Barnett 和 T. Lewis 在文献[6]中说：当所有的工作已经完成后，在异常值研究的重要问题中，只剩下一个最简单的——什么是异常值.

在极值统计研究的问题中，首先是建立一个极值的数学模型. 如果已知观测数据所服从的分布(称之为底分布)，就可以分别得到最大值及最小值的精确模型. 但在大多数应用中，观测数据所服从的分布是未知的，因此只能得到极值的渐近分布，而不是精确分布. 在应用中，这相应于要求数据有比较大的规模. 而且实际证明，在大多数情况下极值的渐近分布提供了一个简单、满意的模型，建立在此基础上的统计分析方法也已得到肯定. 极值统计理论就是为观测到

的基于某个样本量的极值建立一个概率模型，但必须具备某些基本条件：(1)观测对象是随机变量；(2)这个随机变量的底分布应保持不变，或者如果有任何变化，应该可以经数据变换减少这种变化带来的影响；(3)观测到的极值(不是观测数据本身)是独立的，否则需对模型进行相应的修正.

极值统计分析方法区别于一般统计方法主要在于数据的收集，而不是数据的分析. 首先要收集到有资格被称为极值的观测数据，它们满足上述三个基本条件，且有一定的数量规模. 不同建模方法对数据有不同要求：对极值的经典模型，即规范化样本最大值的渐近分布模型，只有“年最大值”或“区组最大值”才可以作为极值的观测数据；而 r 个最大次序统计量模型中，则可将每年或每个区组内的 r 个最大值都作为极值的观测数据；对阈值模型，如果数据是独立同分布的，超阈值近似服从广义 Pareto 分布，而对平稳时间序列，只有超过阈值的峰(POT)可以作为极值的观测数据；对点过程模型，落在远离原点区域上的点组成非齐 Poisson 过程，对相应的似然函数都有一定的贡献.

从尽可能多地利用包含在数据中的信息来看，后面的几种模型都比经典模型好，但问题在于对阈值模型如何确定阈值，几乎等价地在 r 个最大次序统计量模型中，如何确定 r；对点过程模型，如何确定远离原点的区域. 选取阈值是否合适，关系到极值理论应用的成败，即是否能得到一个合理的推断. 如果选取了过高的阈值，将使得手中的数据几乎都处在阈值以下，难以发挥应有的作用，只有极少几个比所选阈值大的数据，才能用于极值的统计分析，造成信息浪费，而且数据太少，结论也不够稳定. 而过低的阈值，又不符合极值模型的理论要求. 因此对一个实际问题，需要在这二者之间进行适当的平衡，既尽可能充分地利用包含在数据中的信息，又能在适当水平上保持模型的正确性. 本书介绍的几种阈值选取法，就是这样一种艺术. 犹如“陶吧”里的一堆土，经艺术家手中出来的是一件精致的陶艺品，而常人则或许做不出一件有用的器皿.

1.2 极值统计的历史

在统计学发展历史中，统计学家首先注意到的自然是随机变量可能取值的主体，不会立即去关心稀有事件，因此极值统计发展的历史相对较短. 历史上，最早可追溯到 1709 年 Nicolaus Bernoulli 讨论的一个精算问题：n 个同龄人在 t 年内死亡，那么平均说来，最长寿者的年龄是多少？他将这个问题简化为一条长度为 t 的直线上的 n 个随机点，离原点的平均最大距离是多少.

在统计文献中，最早讨论极值是 1824 年 J. B. J. Fourier 的一篇文章，他认为与正态分布均值偏离了二个标准差的平方根的三倍的概率大约为五万分之一，即 $P\{|X-\mu|>3\sqrt{2\sigma}\}\approx 1/50\,000$，因此可能完全忽略这类观测. 类似地，按通常的 3σ 原则，认为正态样本的有效范围应在离均值正负三个标准差内. 实际上，这些说法都不够完善. 1877 年 Helmert 指出，这类问题的正确提法应该与样本量有关. 因为当样本量趋于无穷时，有更多的机会使样本最大值出现在分布的尾部，正态总体的样本最大值也应该趋于无穷. 因此，从理论上说，样本最大值与总体均值的距离大于任一固定常数的事件终究要发生. 3σ 原则对小样本来说，有点保守；而对大样本，又太宽松. 极值理论就是说明极值大小与样本量之间关系的理论.

极值的近代理论开始于德国. 1922 年，L. von Bortkiewicz 研究了正态分布的样本极

差[11]，这个问题的意义在于告诉大家，来自正态分布的样本最大值是一个新的随机变量，具有新的分布，因此 Bortkiewicz 是第一个明确提出极值问题的统计学家. 1923 年，德国的 R. von Mises 研究了样本最大值的期望[129]，这是研究正态样本极值的渐近分布的开始. 极值理论的真正发展是 E. L. Dodd 在同年的工作，他首先研究了一般分布的样本最大值[50]. 最重要的结果是 1925 年 L. H. C. Tippet 的正态总体各种样本量的最大值及相应概率表、样本平均极差表[197]. 1927 年，M. Fréchet 发表了第一篇关于最大值的渐近分布的论文[67]，指出来自不同分布，但有某种共同性质的最大值可以有相同的渐近分布，还提出了最大值稳定原理. 但他的文章发表在波兰 Krakow 出版的一份期刊上，而且底分布的类型不是很常用，因而没有得到应有的重视. 1928 年，R. A. Fisher 与 L. H. C. Tippet 发表的文章[65]，现在认为是极值分布渐近原理的基础，他们不仅与 Fréchet 独立地找到 Fréchet 分布，而且还构造了另外二个渐近分布，即极值类型定理. 在这篇文章中，他们第一次描述了正态样本的最大值分布，指出收敛速度是极其缓慢的，这就是以往研究中遇到困难的原因.

1936 年，R. von Mises 提出了最大次序统计量收敛于极值分布的简单有用的充分条件[130]；1943 年 B. Gnedenko 给出了类型定理的严格证明[75]，建立了严格的极值理论，给出了极端次序统计量收敛的充分必要条件. 最后，由 De Haan 进一步研究了 Gnedenko 的工作，将这些结果联系起来，完全解决了吸引场问题[80][81].

最初，极值的概率理论只是研究独立同分布随机变量的最大值或最小值的渐近性质，后来发展到研究次序统计量的分布性质，再后来，研究由底分布的上尾或下尾部确定的在一个高(低)阈值以上(下)关于底分布的超阈值性质. 反过来，底分布的尾部或参数函数可通过极端次序统计量或超阈值用统计方法进行估计. 将注意力集中在分布的尾部可以引入某些特别适合于尾部的参数模型.

上面提到的只是极值统计理论的发展. 20 世纪 20 年代与 30 年代中期，极值统计在气象、人类寿命、放射性、材料强度、洪水、地震、雨量分析等问题中得到了应用. 第一个将样品强度与极值分布联系起来的是英国棉业协会的 F. T. Peirce[146]. 在应用方面，做出最大贡献的是著名的瑞典物理学家和工程师 W. Weibull，他第一次强调极值概念对描述材料强度的重要性[199][200]. E. J. Gumbel 首先向统计学家与工程技术人员提出，应该将极值理论应用于某些他们曾经用经验方法考虑过的分布，于是用极值理论解释了工程界研究了很久的洪水统计分布，以后又用于其他气象现象及异常观测值的统计问题. Gumbel 对极值分析做出了显著贡献，其中大部分都写在他的著作[77]中，这本书奠定了极值理论的基础，成为随机变量极端变异性的建模工具. 虽然有些统计方法已逐渐被更好的方法代替，但至今仍被认为是一本经典著作. 文献[86]总结了到论文发表时的 1978 年的有关极值理论与应用的几乎所有重要文献，具有重要的意义. 近年来出版过更多关于极值的渐近理论与它们的统计应用的著作，例如文献[38]及文献[2]讨论了次序统计量的渐近理论，文献[117]发展了离散与连续随机过程的极值理论. 文献[69]对这个问题给出了比较详尽的叙述，文献[158]主要研究独立同分布随机变量，给出了极端次序统计量的联合分布，多元极值分布第一次出现在这本书中. 文献[109]介绍了极值理论在各个方面，特别在工程领域的应用. 文献[154]讨论各种与极值及次序统计量有关的收敛概念及收敛速度，文献[7]的重点是在精算方面的应用. 文献[57]、[111]、[157]、[22]各有特色，但讨论极值概率模型的较多，也有一些是关于极值统计分析的. 有关极值分布及其应用的综合性文献是文献[16]. 最新关于极值的渐近理论与它们的统计应用的著作是文献[64]和[8]，国

内出版的有关著作可能只有文献[235],关于极值统计的专著似乎还未见到.

我们很容易在国内外出版的各种期刊上找到 1 000 篇以上有关极值统计的文献,这么多的文献足以说明极值分布以及相应的方法极具活力,有广泛的应用,但也反映了理论研究人员与各个不同领域的专业技术人员之间缺乏应有的合作.国内同样存在这个问题,本书目的就是向非统计专业的管理和技术人员介绍极值统计方法,架起交流沟通的桥梁.

最近几年,极值理论的研究取得了很大的进展,对极值理论感兴趣的,已由最初的概率理论的研究人员及实际应用部门,发展到现在的主流统计学家,理论与应用之间的联系也在不断加强.

在极值统计发展的历史中,如果要评价正态分布在其中的作用,我们认为将正态分布作为研究的出发点阻碍了极值理论的发展,因为没有一个基本极值定理与正态分布有关.将正态分布作为研究目的是合理的,因为正态分布是许多近代统计推断的基础.在极值理论中,指数分布比正态分布更加有用,因为此时的基本定理有比较简单的表示,容易进一步研究,所得结果可以推广到其他分布,特别是指数族中的分布.

近半个多世纪以来,计算机的广泛应用对数理统计,包括极值统计在内,在理论、方法和应用上的发展产生了一定的影响.没有现代计算机,就没有现代的统计应用.许多重要的统计方法的应用,都牵涉到大量计算.一元极值,已成功地在许多领域得到应用,各种方法也比较成熟,大都配有用各种语言(Fortran、SAS、Matlab、S-plus、R)编写的计算程序,使用非常方便,这反过来又促进了理论的发展.我们特别推荐使用 R 语言,因为 R 得到了全世界许多优秀统计计算专家的支持,他们自愿奉献给全世界同行免费使用.

现在已经出现专门用于极值统计分析的软件,除了 R 中的 evir,evd,ismev,evdbayes 外,还有文献[157]中附带的一张光盘,带有 Xtremes 软件包,由 J. Beirlain,L. VanAcker,P. Vynckier 于 1996 年设计的软件 ANEX(A statistical ANalysis program with emphasis on EXtreme values).在风险管理方面,特别关于 VaR 的计算,由 JP Morgan 开发的 RiskMetrics 是一种能够测量不同交易,不同业务部门市场风险的软件.有关 S-plus 的介绍可见文献[139]、[60],用于极值分析的程序则可由 Alexander McNeil 的个人网页 http://www.math.ethz.ch/~mcneil/sofrware.html 或 Lancaster 大学数学与统计学系 Jan Heffernan 的个人网页 http://www.maths.lancs.ac.uk/~currie 下载.有关 GPD 参数的置信区间程序由 Nader Tajvidi 提供 http:// www.math.chalmers.se/~nader/software.html.想了解 R 的读者可阅读文献[147],而用于极值分析的程序可从 Lancaster 大学数学与统计学系 A. Stepheson 的个人网页 http://www.maths.lancs.ac.uk/~stephena/下载.

多元极值理论是建立在一元极值理论基础上的,对其进行比较广泛的研究,使其真正得到重视和发展,不过十几年的时间.因此,专门的统计分析方法并不很多,许多统计方法还处在探讨之中,不很成熟,相应的计算程序也不多见,需要读者自己动手编写.

1.3 极值理论的应用

可以说,极值理论是数学在近代工程、环境及风险管理问题应用中取得最成功的重要例子之一.近 50 年来,极值理论已发展成为应用科学中一种非常重要的统计方法,在许多领域都有广泛的应用.

在金融市场,极端事件本身就非常令人关注.近年来,国际上金融危机不断发生:1987 年出现了较大范围的股市崩盘,1995 年 2 月 26 日具有 233 年悠久历史的英国 Barings 银行宣布破产,美国 Orange 县政府的破产,日本大和银行巨额交易亏损等.特别是 1997 年以来的亚洲金融风暴使许多金融机构陷入困境,对我国也有某些直接影响,国内金融界对金融风险有深刻体会,关于金融风险的研究也正在深入,文献[232]可以说是这方面的一个总结.

风险管理的基础和核心是风险测量,对金融市场,就是研究由于市场因子的不利变化而导致金融资产(证券组合)价值损失的大小.现在 VaR(Value at Risk)及极值理论已经成为主流方法,VaR 是一种能全面测量复杂证券组合的市场风险的方法.简单地说,VaR 的概率意义即是损益分布的分位点,估计处于分布尾部的高分位点正是极值理论的最显著特点.

随着 VaR 作为风险度量指标的广泛应用,也逐渐暴露了它的一些缺点.首先 VaR 只关心发生重大损失的可能性,不能给出发生重大损失时可能损失是多少.另一个问题是 VaR 在数学上不具有次可加性.一个简单的例子是只有二种证券 X_1, X_2 组成的组合 $X_1 + X_2$,应该有

$$\mathrm{VaR}(X_1 + X_2) \leqslant \mathrm{VaR}(X_1) + \mathrm{VaR}(X_2)$$

即在同样条件下,证券组合的损失不应超过各个证券损失之和,这就是次可加性.但上述不等式不一定成立.为改进这个不足,提出条件 VaR(Conditional Value at Risk,简记为 CVaR),也称为期望亏空(Expected Shortfall)[3].

在保险业,对大的自然灾害以及如"9.11"恐怖事件所造成损失的赔付,是任何一家保险公司无法独自承担的,这使再保险研究更有必要,我国于 2003 年底成立了中国再保险集团股份有限公司.近几年,全国范围内气象灾害频繁发生,也就是遇到了自然环境中的某些极端情况或罕见的事情.如 1998 年长江流域的特大洪水,2002 年的全国范围的特大干旱,又如 2003 年春季国内许多地区出现了几十年未遇的暖春.我国沿海每年都有不同程度的风暴潮发生,据有关资料记载,在 1949 年到 1993 年的 45 年中,共发生最大增水超过 1 米有 269 次,平均每年 6 次;发生最大增水超过 2 米有 49 次,平均每年 1 次;发生最大增水超过 3 米有 10 次.在信息技术迅速发展的今天,网络通讯占有越来越重要的地位,网络拥塞成为进一步发展的瓶颈.通讯技术及相应硬件的发展跟不上网络用户及网络通讯量的急剧增长,要提高网络通讯质量,不使信息丢失,需要研究如何从数学理论上描述、分析、解决网络拥塞问题.又如在致癌机理研究中,常常考虑化学药品的最大剂量或最低辐射水平.分析解决这些问题,极值理论大有用武之地.

推断性统计是研究如何有效地从已经得到的受随机性影响的观测数据(随机现象的观测资料)提取出尽可能可靠、精确的信息.极值统计则是研究随机变量,或一个过程的取值特别大或特别小情况的随机性质.极值统计分析要求估计的常常不是已经观测到的一般事件的概率,而是在特殊情况下发生的极端事件的概率.举一个例子,上海市是我国最大的工业、贸易、经济、金融中心,但地处长江入海口,3.2m 以下的低洼地面积占全市面积的五分之一以上,黄浦江流经市中心.上海市又位于东亚季风盛行地区,频受台风侵扰,必需修建防汛墙以抗御洪水对上海市的侵袭.那么防汛墙应修多高比较合适?为此,1963 年上海市城建局首次颁发了市区防洪标准,1990 年 12 月水电部上海勘测设计院编制的"上海市黄浦江综合治理规划报告"通过论证,正式提出上海市区的远景防汛标准应为抗御万年一遇的高水位.这是一个目标,在此目标下,估计黄浦公园和吴淞站的相应水位分别高达 6.38m 及 6.81m.这里,我们关心的不是长江、黄浦江的日常水位,而是汛期的最高水位.吴淞观测站始建于 1912 年,黄浦公园站建

于1915年,实际观测资料都不足一百年.如何由这些历史相对较短的资料去估计未来较长时间内的上海市可能遇到的最高水位,这是极值统计应该研究的问题.

能够从水文学或更一般地从自然科学本身导出极值模型,这是最理想的情况,或者依据过去的观测,已经建立了一个公认的、行之有效的模型,也是可以接受的.如果都没有,那么根据极值理论建立一个渐近模型,应该是比较现实的方法.假定 $X_1, X_2, \cdots$ 为吴淞观测站记录的黄浦江每小时水位的高度,则

$$M_n = \max\{X_1, \cdots, X_n\} \tag{1.1}$$

为 n 个观测期最高水位高度.在 $X_1, X_2, \cdots$ 为独立同分布的随机变量假定下,如果我们知道 X_i 的分布及 n,那么 M_n 的分布就能精确地计算出来,

$$\begin{aligned}\Pr(M_n \leqslant x) &= \Pr(X_1 \leqslant x, \cdots, X_n \leqslant x)\\ &= \Pr(X_1 \leqslant x) \cdots \Pr(X_n \leqslant x)\\ &= F^n(x),\end{aligned}$$

这里,$F(x)$ 是 X_i 的分布函数.但实际上 X_i 的分布并不知道,因此就不可能精确计算 M_n 的分布.然而,在相当广泛的条件下,当 $n \to \infty$ 时,经适当规范化,可以得到 M_n 的渐近分布.对于较大的 n,就用这个渐近分布作为 M_n 分布的近似,称为经典模型,这里包含了一个模型外推原则.当然,对这种方法人们也可能提出反对意见.因为,只容许对较大的样本,将数学上 $n \to \infty$ 时的渐近分布作为 M_n 真实分布的近似,但将这个模型外推到没有观测数据出现的范围,还必须进一步假定每小时水位高度足够光滑,即在考虑的时间范围内,每小时水位高度的分布保持不变.

虽然我们认为经典模型是很好的模型,但同时必须指出它的局限性.首先,模型是根据渐近理论建立起来的,因此对有限样本量,模型不是精确的结果.对于一个实际问题,经典模型提供的永远是一个近似的模型,只要这种近似程度是可以接受的.其次,模型本身是在理想情况下得到,与我们所研究的过程可能不很相似(或不是很合理).如(1.1)中吴淞观测站记录的黄浦江每小时水位高度的独立性假定可能不很合理.实际上,相继的前后两个小时的水位高度具有某种程度的相关性.但如果这种相关并不很强,独立性假定仍然是可以接受的.另一方面,随着全球范围内的气候趋于变暖,导致海面水位逐年缓慢上升,海面水位高度的分布实际上也不能保持不变,因此我们研究的只是理想情况.最后,在实际应用中,这样建立起来的模型可能会浪费某些信息.因为记录极值数据常用的方法是只保留一段时间内,比如一年内的最大值,此时(1.1)中的 n 就是一年的观测个数 24×365,这个 n 足够大了,由渐近理论得到的模型描述了年最大值的分布,可以拟合年最大值的变异性,但是能为我们利用的数据只有 $1/(24 \times 365)$.有时可能会有下述情况发生,某一年的年最大值很大,即使是这一年的次大值,比其他年的年最大值大得很多,但是年次大值不能作为年最大值,因而被排除在研究之外,造成信息的浪费.

选择恰当的极值模型后,统计方法在处理极值问题中就显得尤其重要,我们提出以下几点注意.

1.估计方法 估计即是基于现有的观测数据去估计模型的未知参数.对于极值模型的参数估计,第3章将给出许多种方法,如频率直方图用于估计密度,概率图用于拟合分布,以及其他估计方法,包括极大似然估计、矩估计、Bayes估计等.

关于极值的统计分析,有一些方法是只适用于极端数据的,如针对截尾样本的估计方法,以及极值指标、尾部相关参数等的概念.我们将给出各种不同方法,但没有从理论上比较它们的优良性.如果读者认为某种方法比较适合解决所遇到的某个实际问题,那么可以进一步考察统计分析的结果是否满意,在此基础上再去考虑其他方法.所有的估计方法都有各自的优点和缺点,我们认为极大似然估计法是较好的,主要原因有三:首先,它是唯一能够适应模型变化的方法,对各种不同极值建模方法得到的模型都适用,尽管不同的极值建模方法得到的模型也有不同,极大似然估计量的表示也会改变,但方法的本质没有任何改变;其次,可以把各种各样的有关信息综合到统计推断中去;最后,最重要的是极大似然估计具有优良的大样本性质,能给出估计方法不确定性的度量.

2.不确定性的定量表示 统计分析即是利用现有的观测数据对真实情况给出"最好的猜测".但如果对所研究的真实过程再一次抽样,将得到不同的样本观测值,因而有不同的估计值.所以,估计模型时必须考虑由于样本变异性引起的模型不确定性.在已知模型类型时,模型的不确定性主要体现在模型参数上,标准误是参数的不确定性或它的变异性度量.在极值模型中,模型参数的很小变化可能使外推结论发生很大改变.也就是说,极值问题中可能存在的不确定因素比其他统计问题还多.因此,估计一个过程在极端水平上的不确定性,同水平本身作为一个参数必须进行估计一样重要.遗憾的是,不确定性的度量在应用中还是常常被忽视.我们将看到,由于极大似然估计的渐近正态性,容易给出估计值及其标准误.

3.模型诊断 一个极值模型能够用于实际问题,唯一理由是导出此模型的渐近性.如果已经发现一个模型与观测到的极值拟合得不好,再进行外推也不可能得到好的结果.我们将在以后对每一极值模型都给出几种评价拟合好坏的方法.

4.信息的极大使用 尽管不确定性是任何统计模型所固有的,但是如果能谨慎地选择模型和推断方法,尽可能充分地利用所有的信息,这种不确定性是可以减少的.本书将给出许多可供选择的模型,如不仅有区组最大值模型,还有可利用更多数据的模型,利用协变量所提供的相关信息的模型,以及多元模型,利用附加知识或信息的 Bayes 模型等.这些将在下面各章节中讨论.

2 一元极值理论

本章讨论极值的基本理论，这是本书的基础. 内容包括由经典的极值理论导出的极值分布，由阈值模型导出的广义 Pareto 分布，稳定分布、厚尾分布、次指数分布以及与极值有关的分布. 定理 2.1 给出的 Fisher-Tippett 的极值类型定理是极值分布渐近原理的基础，本章首先给出经过规范化最大值的极限分布形式，然后介绍极值分布的最大值吸引场，给出规范化常数(norming constants)的计算公式.

2.1 经典的极值理论

极值理论曾经是一个引起不少统计学家注意的问题，他们主要研究了以下两个问题：究竟有哪些分布可以作为极值分布；收敛到某个特定的极值分布的条件是什么？极值类型定理回答了前一个问题，后一个问题称之为极值分布的最大值吸引场条件.

2.1.1 极值分布的类型及其性质

设 $X_1, X_2, \cdots$ 是独立同分布的随机变量，分布函数为 $F(x)$(称为底分布)，对自然数 n，令

$$M_n = \max\{X_1, \cdots, X_n\}, \qquad m_n = \min\{X_1, \cdots, X_n\} \tag{2.1}$$

分别表示 n 个随机变量的最大值与最小值，则

$$\Pr(M_n \leqslant x) = \Pr(X_1 \leqslant x, \cdots, X_n \leqslant x) = F^n(x), \quad x \in \mathbb{R},$$

$$\Pr(m_n \leqslant x) = 1 - \Pr(m_n \geqslant x) = 1 - [1 - F(x)]^n, \quad x \in \mathbb{R},$$

这里 $\mathbb{R}$ 表示所有实数集合. 如果已知分布函数 $F(x)$，就可以根据上式，精确地求出最大值和最小值的分布函数. 但在应用中，F 往往是未知的，因此很难直接用于统计分析. 所以，我们需要研究最小值 m_n 和最大值 M_n 的极限分布，它有很重要的理论和实际意义. 若记

$$A = \{x : 0 < F(x) < 1\}, \qquad x^* = \sup_{x \in A} A, \qquad x_* = \inf_{x \in A} A,$$

称集合 A 为分布 F 的支撑，x^* 和 x_* 分别为分布 F 支撑的上端点和下端点. 显然对所有 $x_* \leqslant x < x^*$，都有

$$\Pr(M_n \leqslant x) = F^n(x) \to 0, \qquad n \to \infty.$$

如果 F 的上端点 x^* 有限，即 $x^* < \infty$，则当 $x \geqslant x^*$ 时，有

$$\Pr(M_n \leqslant x) = F^n(x) \to 1, \qquad n \to \infty.$$

这就是说，不论 x 是否有限，当 $n \to \infty$ 时，最大值 M_n 分布的极限只能是 0 或 1，这种退化分布是没有任何意义的，因此我们不直接讨论最大值的渐近分布. 类似于处理 n 个随机变量之和的中心极限定理，我们试图通过对 n 个随机变量最大值 M_n 的规范化变换，以了解最大值分布的性质.

定理 2.1 (Fisher-Tippett 的极值类型定理)设 $X_1, \cdots, X_n$ 是独立同分布的随机变量序

列，如果存在常数列$\{a_n>0\}$和$\{b_n\}$，使得

$$\lim_{n\to\infty}\Pr\left(\frac{M_n-b_n}{a_n}\leqslant x\right)=H(x),\quad x\in\mathbb{R} \tag{2.2}$$

成立，其中$H(x)$是非退化的分布函数，那么H必属于下列三种类型之一：

Ⅰ型分布：$H_1(x)=\exp\{-e^{-x}\},\quad -\infty<x<+\infty$；

Ⅱ型分布：$H_2(x;\alpha)=\begin{cases}0, & x\leqslant 0,\\ \exp\{-x^{-\alpha}\}, & x>0,\end{cases}\quad \alpha>0$；

Ⅲ型分布：$H_3(x;\alpha)=\begin{cases}\exp\{-(-x)^{\alpha}\}, & x\leqslant 0,\\ 1, & x>0,\end{cases}\quad \alpha>0.$

其中Ⅰ型分布称为Gumbel分布，Ⅱ型分布称为Fréchet分布，Ⅲ型分布称为Weibull分布，这三种分布统称为极值分布(extreme value distribution)．当$\alpha=1$时，$H_2(x;1)$，$H_3(x;1)$分别称为标准Fréchet分布与标准Weibull分布．称a_n，b_n为规范化常数．■

极值类型定理说明，如果M_n经线性变换后，对应的规范化变量$M_n^*=(M_n-b_n)/a_n$依分布收敛于某一非退化分布，那么，不论底分布$F(x)$是何种形式，这个极限分布必定属于极值分布的三种类型之一．因此，极值类型定理提供了类似于中心极限定理的极值收敛定理．证明参见文献[117]．

从模型的角度来看，三种极值分布类型$H_1(x)$，$H_2(x;\alpha)$和$H_3(x;\alpha)$完全不同，但从数学的角度来看，它们之间却存在着非常密切的关系．事实上，可以直接验证下面的结论：设$X>0$，则

$$X\sim H_2\Leftrightarrow \log X^{\alpha}\sim H_1\Leftrightarrow -X^{-1}\sim H_3.$$

因此在某些场合，为方便起见，可以假定其中任意类型的极值分布．

极值分布的最大值稳定性

定义 2.1　对于给定的分布函数$F(x)$，如果存在序列$\{a_n>0\}$，$\{b_n\}$，使得

$$F^n(a_nx+b_n)=F(x),$$

则称分布函数$F(x)$是最大值稳定的(max-stable)．▲

由(2.2)知，若$F(x)$是最大值稳定的，则相应的M_n的分布仍然是$F(x)$．对于极值Ⅰ型分布，取$a_n=1$，$b_n=\log n$，不难验证

$$H_1^n(x+\log n)=H_1(x).$$

所以，极值Ⅰ型分布是最大值稳定分布．

同理，对于极值Ⅱ型和Ⅲ型分布，分别取$a_n=n^{1/\alpha}$，$b_n=0$和$a_n=n^{-1/\alpha}$，$b_n=0$，有

$$H_2^n(n^{1/\alpha}x;\alpha)=H_2(x;\alpha),\qquad H_3^n(n^{-1/\alpha}x;\alpha)=H_3(x;\alpha).$$

所以，极值Ⅱ型分布和极值Ⅲ型分布也都是最大值稳定分布．

事实上，有进一步的结论：一个分布函数$F(x)$是最大值稳定分布，当且仅当$F(x)$是三种极值分布之一(证明参见文献[57]定理3.2.2)．

极值分布的密度函数　容易求得三种类型极值分布的密度函数分别为

$$h_1(x)=e^{-x}H_1(x),\quad -\infty<x<+\infty;$$

$$h_2(x;\alpha)=\alpha x^{-(1+\alpha)}H_2(x;\alpha),\quad x>0;$$

$$h_3(x;\alpha)=\alpha(-x)^{\alpha-1}H_3(x;\alpha),\quad x\leqslant 0.$$