黄幼才　编著

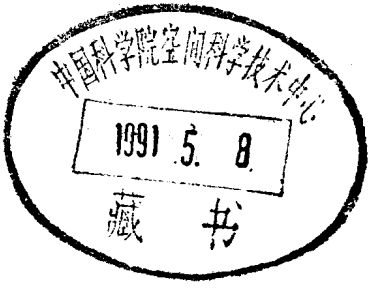
数据探测与抗差估计

测绘出版社

P207
HYC

数据探测与抗差估计

黄幼才　编著



测绘出版社

106767

本书综合了国外大量文献和作者本人研究成果编写而成。全书共分十二章。前三章介绍了数据探测，后九章介绍了抗差估计理论。其中主要包括有抗差性度量，抗差估计的设计，常用的几种M估计及其计算方法，多维M估计，方差和协方差阵估计，回归分析等。

可供数学工作者、测绘、医学、化学、经济学和其他工程部门的有关业务技术人员，有关院校师生和科研人员参考。

数据探测与抗差估计

黄幼才　编著

*

测绘出版社出版

涿州市冶林印刷厂印刷

新华书店总店科技发行所发行

*

开本 850×1168 1/32 · 印张12.375 · 字数·274千字

1990年11月第一版 · 1990年11月第一次印刷

印数 0,001—1,100册 · 定价 15.00元

ISBN 7-5030-0404-5/P · 142

前　言

如何有效地消除或削弱观测数据中的粗差对参数估值的不良影响，一直是人们关注的问题。经典的估计理论把观测值的分布局限在某种理想的分布模式，一旦观测数据偏离了这种分布模式，参数估值的效率和可靠性就会严重地受到损害。尽管人们采取多种方法来排除粗差的干扰，但是由于这些方法几乎不可能完全把参数估值的计算和排除粗差分开，因而这些方法的可靠性和排差能力也受到限制。近四十年来，一些统计学家一直致力于发展一种新的估计理论，旨在寻找一些既能抗拒粗差影响又基本上具备经典最优估计的一些优良的统计特性的估计，这就是本书讲的抗差(Robust)估计理论(或称稳健估计)。所谓抗差是指参数估值能抵御观测数据中大量存在的小误差和少数粗差的干扰。抗差估计对观测值的分布要求不十分严格，它只要求子样“近似”地服从某一分布模式。这种“松散”的假设，使得抗差估计能避免由于模式分布的变化所带来的不良影响，因而也就削弱和消除了粗差对参数估值的影响。事实上，完全符合某一特定分布的观测数据是不可能得到的。这也就说明了抗差估计是一种更切合于实际的估计理论。

虽然很早就有人提出抗差估计，但是直到60年代，抗差估计理论的重要性才广泛地为人们所认识。抗差估计理论上的系统化和完善还是近几年的事。我国许多学者在推广和运用这种理论方面做了许多工作。但是，有关系统阐述抗差估计理论的中文书籍极少，这无疑给这种理论的推广和深化带来一定的困难。作者曾在国外从事这方面的理论学习和研究，此书综合了国外大量文献和作者本人研究成果而写成。希望这本书能起一种抛砖引玉的

作用。

本书是以工科院校研究生、高年级本科生和从事数据处理的同志为对象编写的。因此不太追求数学上的系统与严谨。为了便于读者对这一理论的理解，本书加强了经典估计理论与抗差估计理论的衔接，侧重于实例中的公式推演和分析。

本书是在周江文研究员的鼓励下写成的。书稿完成后，周先生和陶本藻教授在审阅中提出了许多宝贵意见，在此表示衷心的感谢。

由于作者的水平有限，书中出现一些错误和不足之处在所难免，作者期待广大读者提出宝贵意见。

黄幼才

1989 年

此为试读，需要完整PDF请访问：www.ertongbook.com

目　录

绪　论

传统的估计理论建立在随机变量的母体严格服从某一特定的分布模式这种假设基础上。从母体中随机取得的 n 个值称为“子样”，n 称为子样的“容量”。因为子样与母体有相同的分布，所以可以利用子样对母体的统计性质作出推断。由子样可以得到一个离散分布，这种子样分布叫做经验分布，其对应的母体分布称为理论分布。当子样容量 n 足够大时，可以用经验分布代替理论分布。实际上，是通过一组观测值来求定母体的参数。观测本身可以理解为抽样，一组观测值可以认为是从母体中随机取得的一个子样。由于观测值的个数 n 是有限的，因此不可能得到参数的真正数值，而只能对它们作出“估计”。这种用有限个数的观测值来估计参数的问题称为“参数估计”。例如，最小二乘估计是利用一组来源于母体为正态分布的观测值来求定母体参数的一种参数估计。它的最优统计特性不能脱离观测值的母体是正态分布这一假设前提。

然而，实际上几乎不可能找到一个完全符合于正态分布的母体。测量仪器不完善，观测中不可避免的错误等因素都可能使观测值偏离正态分布。此外，观测值也有可能来源于其它类型分布的母体。大量的实验和研究表明，当观测值的分布偏离了原来的假设模式时，由经典的数理统计方法所得的参数估值往往很不可靠。例如在一维正态分布的数据中，一个粗差会彻底破坏最小二乘估值，因为每一个观测值都对参数的最小二乘估值产生影响。近三十年来，统计学家们通过大量的实验和理论研究一直寻求一些能抗拒粗差影响的估计。数据探测技术和抗差估计理论就是在

这种背景下提出和发展起来的。

数据探测技术能使我们以一种合理的推理方法来分析数据，能使我们较快而且容易地勾画出数据的基本结构。为了得到可靠的参数估值，在进行统计分析和检验之前要仔细观察和分析数据的结构，然后采用适合于这种数据的估计。数据探测的特点是直观，不需要很深的数学知识。

数据探测主要包括三部分：图表分析，数据变换，余差分析。图表分析是利用作图的方法把数据的主干和随机部分分离出来。从图中我们可以直观地了解数据的结构、分布图形、对称情况、离异值等。余差分析是一种传统的方法。这里所指的余差是指采用抗差估计后所得的余差，即排除了少数粗差的干扰。通过余差分析我们基本上对误差分布的对称性、尾长、离散情况等有所了解，以证实我们采用的估计的合理性。数据变换是对原有观测数据进行数学变换。对数据变换的要求是保持原有数据的基本结构不变，即原有观测值排列的次序不变，中心位置不变，不改变原有数据分布曲线的平滑性等。数据变换的特点是改变了原有数据之间的间隔。数据变换后，观测值的尺度不一致性得到缓和，这有利于定位参数和尺度参数的估计（许多定位参数估计要求尺度不变）。另外，经过数据变换，数据分布更趋于对称，这对定位参数估计是非常有利的，因为几乎所有一维估计的定位参数就是对称中心。例如，算术平均值、中位数、α截尾均值等。在分布模式是对称的情况下，上述几种定位参数估计都是无偏的。数据探测技术只能解决简单的数据分析，对于较为复杂的数据则需要用抗差估计理论。数据探测包含一些基本的抗差概念，它有助于我们对抗差估计理论的理解。

抗差估计的设想很早就有人提出，但是真正引起人们重视还是在60年代，因为电子计算机的出现为抗差估计提供了计算手段，使之能在实际中广泛应用。Reter J.Huber于1964年发表了一篇题为《定位参数的抗差估计》的论文，为抗差估计理论化

奠定了基础。抗差估计理论系统化还是近几年的事。Frank R.Hampel，John W.Tukey 等人为抗差估计理论的发展作出了重大贡献。

抗差估计理论是在经典的统计理论基础上发展起来的。它的指导思想是寻求一些既能抗拒粗差影响，同时又基本上具备传统估计方法的一些优良统计特性的估计。由于抗差估计对子样的母体分布没有严格的要求，因此根据各种不同的分布模式设计出来的估计也不相同，迄今为止，抗差估计大约有近百种。由于分布模式的不唯一，所以不可能找出一种适合各种情况的估计，这给抗差估计理论研究带来一定的困难。

抗差估计虽然对子样的母体分布没有严格的要求，但并不是没有要求。母体分布完全不可知将会给理论研究带来极大的困难。从前面的讨论可知，问题的关键是假设的分布模式是否符合客观实际情况。例如，在研究抗差估计理论时，我们经常假设观测值是来源于受了污染的正态分布母体。它的主体是正态分布，次要部分是方差较大的正态分布或某种对称分布（由粗差引起的）。这两种类型的观测值均值相同，方差不同。两种分布的叠加就是所谓受了污染的正态分布。这种分布模式基本上符合实际情况。正态分布受粗差的影响往往会使其尾部变长。因此，为了研究分布模式污染的程度对估计量的影响，可将观测数据的分布按尾部长短分类，用计算机来模拟各种不同类型的数据，然后用数值来评估各种估计的抗差能力、效率等。这就是所谓 Monte Carlo 实验。美国普林斯顿大学抗差估计研究实验室专门从事这项工作。Monte Carlo 实验是抗差估计理论研究的一个重要组成部分。

估计理论的核心是寻求参数的最优估值。例如，算术平均值是正态分布数据的定位参数的最优估值。当观测数据中出现了0.2%的粗差时，最小二乘估计就失去了它的最优性。粗差的概率为0.2%在实际中是完全正常的。因此由经典估计理论所得估

此为试读，需要完整PDF请访问：www.ertongbook.com

值的最优特性局限在狭隘的范围内。而抗差估计是一种广义的最优估计，它或许对某一特定的分布模式不是最优，但它能排除粗差的干扰，得到能代表数据主体的可靠估值，它的效率接近于经典估计的效率。但是，如果子样的母体分布确定是正态的话，则抗差估计的效率就不如最小二乘估计，因为为了保证估计的抗差性，我们必须要去掉或限制一部分观测值。所以设计抗差估计的指导思想是在抗差能力和效率中求得最佳平衡。要求两者都达到最佳是不可能的，因为效率和抗差能力是相互矛盾的两个方面。

抗差估计基本上分三大类：广义极大似然估计(M 估计)，排序线性组合估计(L 估计)，列序估计(R 估计)。M 估计是经典的极大似然估计的推广。相对其它两种估计，M 估计比较容易地过渡到抗差回归估计、抗差方差和协方差估计等。所以本书主要讨论了 M 估计。

全书共分十二章。前三章介绍了数据探测原理，引进了抗差估计理论中的一些基本概念。后九章系统地介绍了抗差估计理论。其中包括抗差估计的设计，常用的几种 M 估计，M 估计的计算方法，方差和协方差估计，回归分析，方差分析等。另外，用几个测量方面的实例讨论了抗差估计理论的应用。最后附录中详细推演了方差变化函数，并列出了本书中将要涉及到的一些数理统计和集合学中的基本概念。

第一章　图表分析

一般来说，数据的分布结构很难直接通过对数字本身的研究来发现。用图表的方法重新安排这些数据能把我们的注意力从数字引向数字各种不同的特点，从而发现其内在规律。利用图表分析我们可以观察到：

（1）数据的对称性如何；

（2）数据的离散情况；

（3）是否有少数数字偏离主体；

（4）整个数据是否集中在一部分；

（5）数字之间的间隔情况。

第一节　枝叶图

枝叶图法是把数据按其大小排列，然后找出数据主干和次要部分。主干部分称为枝，次要部分称为叶。在使用这种方法时主要考虑的是区间数和区间的宽度。现用实例来说明这种方法。

表 1.1 列出了某医院对 21 人进行体温与身体某一生理特征周期之间关系的研究结果，其中体温记录数据正常。显然，从数字本身看不出其规律性。

现在用枝叶图来探测表 1.1 中的数据分布结构。首先讨论周期的枝叶图，表 1.1 中的整数部分定为枝，小数部分定为叶。没有叶则枝也不存在。图 1.1 是表 1.1 中周期的枝叶图。

图 1.1 表明有 2/3 的人的平均周期在 26.3 和 28.8 之间，其中 22.9 被看成是不正常的离异值。

深度枝叶(单位=0.1天)

	22	9
	23	
	24	
	25	
6	26	3 6 8 9 9
9	27	5 6 6
(6)	28	0 4 4 5 8 8
6	29	4 9
4	30	0 3
2	31	2 8

图 1.1 周期枝叶图

21 人体温与周期关系 表 1.1

序 号	周 期(天)	温 度(°C)
1	22.9	36.44
2	26.3	36.21
3	26.6	36.71
4	26.8	36.13
5	26.9	36.25
6	26.9	36.53
7	27.5	36.41
8	27.6	36.45
9	27.6	36.53
10	28.0	36.31
11	28.4	36.63
12	28.4	36.54
13	28.5	36.52
14	28.8	36.62
15	28.8	36.40
16	29.4	36.48
17	29.9	36.39
18	30.0	36.37
19	30.3	36.77
20	31.2	36.76
21	31.8	36.50

图 1.1 的第一列是数据的深度。它反映了数据的离散情况。现以数据 26.3 为例来说明数据深度的含义。从图 1.1 可以看出数据是按大小顺序从上往下排列。最小的数是 22.9，最大的数是 31.8。从 22.9 往下数，26.3 的序号是 2。从 31.8 往上数，26.3 的序号是 20。取 2 和 20 两个数中的较小的一个数为 26.3 的深度，即 26.3 的深度是 2 。但是这样定义深度也有个缺点。对于大子样来说，每个数据有一个深度，结果过多的深度反而说明不了数据的离散情况，因此，以数据枝叶图的每一行（每条枝）数

据中最大的深度定义为这行中所有数据的深度，于是，26.3 的深度应为 6 。另外，数据的最大深度不能超过数据总数的一半。这样定义数据深度必然会出现这样一个问题：位于数据中间的一行的深度超过数据总数的一半。于是定义这一行为数据的“中线”，并用括号表示。括号内的数字就是中线包含的数据的个数。显然，中线包含有数据的中位数（简单而言，中位数是数据按大小顺序排列之后的中间的一个数，严格的数学定义后面将给出）。可以用括号内的数字与它上下邻近的两个深度之和等于数据的总数来检查深度的计算是否正确。以图 1.1 为例，9+6+6=21。深度值小说明数据不可靠。

枝叶图能否有效地反映数据的结构主要取决于数据每行宽度的选择。数据所需最大的行数可以由下面的经验公式来计算：

$$L=10\times\log_{10}n$$

其中 n 是数据个数，L 是最大的行数。这个公式对于子样大小为 $20\leqslant n\leqslant 300$ 是有效的。根据表 1.1，$n=21$，则最大行数为

$$L=10\times\log_{10}21\approx 13$$

数据之间的间隔可用下面的方法来计算：

整个区间的长度　$R=31.8-22.9=8.9$

每行的宽度等于　$R/L=8.9/13=0.68$

对计算结果四舍五入求得宽度的最后值是 $R/L\approx 1$ 。因此图 1.1 确定数据整数部分为枝是合理的。

为了排除离异值对枝叶图的影响，应该把这些点从数据主体中分离出来，图 1.2 表示排除数值 22.9 后的枝叶图。

除了直接对数据本身作枝叶图分析外，还可以对回归拟合后的余差作枝叶图分析。如果拟合函数选择正确，观测值中没有系统误差和粗差存在，则余差的结构应是以零为对称中心，中部的余差数目多于两尾的余差数目的对称分布。现以表 1.1 中的数据为例，设因变量 y 为体温，自变量 x 为周期长。经过抗差化直线拟合后得直线方程为

深度 ($n=21$)	（单位=0.1天）
1	低（22.9）
6	26 | 3 6 8 9 9
9	27 | 5 6 6
(6)	28 | 0 4 4 5 8 8
6	29 | 4 9
4	30 | 0 3
2	31 | 2 8

图 1.2 周期枝叶图(排除 22.9 的影响)

深度 ($n=21$)		（单位=0.01℃）
1	−3	0
2	−2	0
5	−1	3 5 8
(6)	0	0 2 4 7 8 9
10	0	3 6 7 9
6	1	1 2 5
3	2	0 3 8

图 1.3 余差枝叶图

$$\hat{y} = 0.02813x + 35.68$$

通过数据枝叶图分析可知，22.9 是一个离异值（或称粗差）。抗差化直线拟合能排除该观测值的影响（抗差化直线拟合将在第三章中详细讨论）。图 1.3 是利用上面抗差化直线方程对体温和周期进行回归拟合所得的余差枝叶图，可以看出余差大部分集中在零的周围，正的余差多于负的余差。

第二节 分布数

枝叶图从定性的角度概括了数据的结构。经典的估计理论是通过几个特征统计量来描述数据结构的。例如，最小二乘估计中的算术平均值和方差用来描述正态分布数据的结构。类似地，在数据探测中是用中位数(M)、四分位数(F)、四分位数离散度(dF)等特征统计量来描述数据结构的。四分位数离散度 dF 是上四分位数（F_U）与下四分位数（F_L）之差。分布在数据两个端

点的数值分别称为上极值和下极值。这些特殊的分位数（即 $1/2^k$ 分位数，取 $k=0$ 或 $k=n$ 得下极值或上极值，取 $k=1$ 得中位数，取 $k=2$ 得四分位数，取 $k=3$ 得八分位数等等）和由这些分位数定义的分位数离散度统称为分布数。需要指出的是，在计算分布数之前，原数据必须按大小顺序重新排列。当子样容量 n 为有限值时，称为经验分位数，它在分布位置上与分位数有微小差别。为了叙述方便，下面的讨论都用分位数。分位数的数学定义将在第七章中给出。图 1.4 是分布数示意图。

图 1.4　分布数示意图

从图 1.4 中可以直观地看出中位数是观测值按大小顺序排列后中间的观测值。四分位数是中位数和极值中间的一个观测值。数据探测技术是用中位数作为定位参数的估值，用四位数离散度作为尺度参数的估值。很显然，M 和 dF 只与观测值的分布位置有关，与极值大小关系不大，也就是与粗差的大小关系不大。因此用分布数来估算定位参数和尺度参数是抗差的。

下面列举一些分布数和对应的尾部面积：

分　布　数	尾　部　面　积
M（中位数）	$\frac{1}{2}=0.5$
F（四分位数）	$\frac{1}{4}=0.25$
E（八分位数）	$\frac{1}{8}=0.125$
D（十六分位数）	$\frac{1}{16}=0.0625$
⋮	⋮

此为试读，需要完整PDF请访问：www.ertongbook.com