# Theory of Point Estimation

Second Edition

# 点估计理论

第二版

E.L.Lehmann
George Casella 著

郑忠国 蒋建成 童行伟 译

中国统计出版社
China Statistics Press

# Theory of Point Estimation

Second Edition

# 点估计理论

第二版

E.L. Lehmann
George Casella 著

郑忠国 蒋建成 童行伟 译

中国统计出版社
China Statistics Press

(京) 新登字 041 号

**图书在版编目 (CIP) 数据**

点估计理论 (第 2 版)/(美) 莱曼-(Lehmann, E.L.),
(美) 卡塞拉 (Casella, G.) 著; 郑忠国等译.
— 北京: 中国统计出版社, 2004.11
ISBN 7-5037-4557-6

I. 点... II. ①莱... ②卡... ③郑... III. 点估计
IV. 0211.67

中国版本图书馆 CIP 数据核字 (2004) 第 108781 号



译　　者 / 郑忠国　蒋建成　童行伟
责任编辑 / 徐　颖
封面设计 / 艺编广告 · 张　冰
出版发行 / 中国统计出版社
通信地址 / 北京市三里河月坛南街 75 号　邮政编码 /100826
办公地址 / 北京市丰台区西三环南路甲 6 号
电　　话 / (010)63459084　63266600-22500(发行部)
印　　刷 / 河北天普润印刷厂
经　　销 / 新华书店
开　　本 / 787×1092mm　1/16
字　　数 / 660 千字
印　　张 / 32
印　　数 / 1—3000 册
版　　别 / 2005 年 8 月第 1 版
版　　次 / 2005 年 8 月第 1 次印刷
书　　号 / ISBN 7-5037-4557-6/O·53
定　　价 / 45.00 元

To our children

Stephen, Barbara, and Fia
ELL

Benjamin and Sarah
GC

# 作者为中译本写的前言

本书第一版于 1983 年出版，其风格上强调理论与应用并重。在第二版 (1998 版) 中，在内容上除了对第一版材料加以常规的更新和充实，特别大大地增加了关于 Bayes 推断和压缩估计的论述，我们特别欣慰地看到中译本的出版将使本书内容在中国统计学界得到广泛了解。我们非常感谢郑忠国教授等在翻译出版过程中付出的辛勤劳动并很高兴看到本书中译本的出版。

本书作者之一， E. L. Lehmann 借此机会向伟大的中国统计学家许宝騄教授表达感激的心情。许宝騄教授在访问 UC Berkeley 期间协助 Neyman 教授指导我的博士学位论文，帮我选题并提供了解决问题的思路。我们也非常愉快地回忆到我和我的妻子 (Juliet Shaffer) 在 1986 年访问北京时受到的令人难以忘怀的盛情接待。

近年来，许多中国学生在美国学习，不少访问学者来美进行学习交流。对于本书的两位作者来说，与中国统计学者的交流是一种享受。我们希望，本书中译本的出版将进一步促进和加强这种国际交流。

E. L. Lehmann
Berkeley, California

George Casella
Gainesville, Florida

2003 年 8 月

# 译者序

本书是作者在收集了大量丰富的素材和积累了多年的教学经验的基础上写成的，是一本很好的教科书，可作为研究生教材。本书有两大特点：第一，尽量避免繁冗的数学推导，力图把统计的思想与方法传授给学生。本书行文通俗易懂，读者只要具有数学分析和高等代数的知识，就能阅读全书。第二，在保持行文通俗流畅的同时，作者注意论述近代估计理论的最新成果。对于即将在统计理论或应用领域进行科学研究的学生来说，这是一本很好的入门书。

鉴于本书的以上两个特点，译者认为本书比较适合我国大学教学的现状，可以作为统计系和数学系概率统计专业研究生及高年级大学生的教材或参考书。对于想了解概率统计方法的实际工作者，本书也不失为有价值的参考书。

在此，我们要特别提及本书作者之一 E. L. Lehmann 教授和中国统计学的渊源。Lehmann 教授曾受到许宝騄教授的帮助，这一点使他难以忘怀。中国学者在与他接触中，只要一提到许宝騄教授，他立刻会兴奋起来，并且说："那我们都是许先生的弟子了"。他对于赴美留学或访问的中国学者，总是热情关怀，觉得这是对许宝騄教授帮助的回报。

本书翻译过程中得到两位作者的热情关怀，同时也得到了中国统计出版社的支持与鼓励。我们的同事陈家鼎教授对于本书的翻译和出版也给予极大的支持。许多研究生主动协助核对译文，他们是李季轩、刘建军、王述珍和赵慧。对于他们的帮助在此一并表示感谢。

译文有错误或不妥之处，欢迎读者批评指正。

译　者

2003 年夏

# 第二版前言

自 1983 年本书第一版问世以来，统计学得到飞速发展，这引发了出版第二版的愿望。新版中由于加入了新材料，内容由 500 页增至 600 页，文中引述的将近 1000 篇参考文献中约 25% 是 1983 年以后出现的。

在第一版中关于 Bayes 推断的论述较少，此次再版中作了很大改变，增加了同变 Bayes 估计、多层 Bayes 估计和经验 Bayes 估计以及它们的比较。此外，还增加了关于信息不等式、同时估计和压缩估计的新近进展的内容。在每章后面的注记中，除了提供文献的历史材料外，还引进了点估计理论和相关主题的最新发展情况。由于篇幅原因，这些新进展不可能包括到本书的正文中去，此外，习题部分也作了很大的扩充。另一方面，由于篇幅原因，我们略去了第一版中关于稳健估计 (特别是 L 估计、 M 估计和 R 估计) 的论述，这是由于近年来已经出版了两本关于稳健估计的专著 (Hampel et. al. (1986) 和 Staudle & Sheather (1990)) 。此外，本书编写体例方面也作了小的修改，例如所有的参考文献都集中放在本书的最后，将本书的例子列成表；对于书中公式或方程的标号，在本章内用节号、公式号表示，例如 (3.2) 代表第三节第二个公式，涉及其它章时，同时列出章号，例如 (3.3.2) 代表第三章第三节第二个公式。

本版对学生具备的统计知识与第一版相同，学生如果具备了掌握 Bickel & Doksum (1977) 或 Casella & Berger (1990) 等教材的统计知识就足够了。第二版的点估计理论 (TPE2) 和第二版的假设检验 (TSH2) 构成姐妹篇，它们共同形成了古典统计的基础教材。

许多同仁对 TPE2 的出版作出了贡献，他们提出了建议，进行审阅甚至给出题解。特别感谢 John Kimmel 对整个出版过程的关注。下列同事对本书进行审阅或给出题解： Matt Briggs , Lynn Eberly, Rich Lerine 和 Sam Wu 。下列同事对于本书的许多主题提出了建议并进行了讨论：Larry Brown , Anirban DasGupta, Persi Diaconis, Tom DiCiccio, Roger Farrell, Leslaw Gajek, Jim Hobert, Chuck Meculloch, Elias Moreno, Christian Robert, Andrew Rukhin, Bill Strawderman 和 Larry Wasserman 。 June Magermann 将本书内容打印成 Latex 文件， Andy Scherrer 协助挽救了由于硬盘毁坏而几乎丢失的文件， Marty Wells 提供了本书写作过程中所需的参考文献。对于以上所列同事的协助，在此一并表示感谢。

E. L. Lehmann　　George Casella
Berkeley, California　　Ithaca, New York

1998 年 3 月

# 第一版前言

本书论述欧氏样本空间中的点估计问题。前四章涉及精确 (小样本) 理论，其写作方法与材料编排类似于其姐妹篇 "统计假设检验 " (TSH) 。在这一部分中根据诸如无偏性、同变性和极小极大性等准则得出优良估计，并围绕这些准则组织材料。这些准则主要应用于指数族和群族。归入这些族的 (相对简单的) 统计问题的内容是很丰富的，系统地讨论这些问题构成了本书的第二个主题。

由于采用大样本方法我们可以得到具有广泛应用价值的理论，因而后两章专门叙述大样本理论。第五章就渐近性概念和方法提出了一个相当基本的导论。第六章在充分正则的条件下，建立了极大似然估计及有关估计和 Bayes 估计的渐近有效性，并简短地介绍了 Hajek 和 LeCam 的局部渐近最优性理论。然而，即使在后两章中，我们的注意力仍限于欧氏样本空间。因此，本书不包括特别和序贯分析、随机过程、函数空间中的估计问题。

本书附有许多习题。它们和参考文献都集中放在各章末尾。关于参考文献，特别是包括了应用问题的文献，其数量是如此巨大，散布在如此众多的国家和专业的杂志上，以至于完全罗列似乎是不可能的。因此这里给出的文献是不太完全的，它们部分地反映了我个人的兴趣和阅历。

本书通篇假定读者具有良好的数学分析和线性代数知识。在接受下述约定下，阅读本书的大部分内容不需要更进一步的数学知识 (包括为了使内容完备而在 1.2 节给出的测度论概述) 。

1. 一个主要的概念是关于如 $\int f dP$ 或 $\int f d\mu$ 的积分，它包括了离散和连续两种情况。在离散的情况下，$\int f dP$ 为 $\Sigma f(x_i)P(x_i)$ ，这里 $P(x_i) = P\{X = x_i\}$ ；而 $\int f d\mu$ 为 $\Sigma f(x_i)$ 。在连续的情况下，$\int f dP$ 和 $\int f d\mu$ 分别为 $\int f(x)p(x)dx$ 和 $\int f(x)dx$ 。作这样的替换之后可以毫无困难地阅读全书 ( 除去记号的统一和概念的一般性外 ) 而不会失去什么东西。

2. 当我们指定一个概率分布 $P$ 时，不仅必须指定样本空间 $\mathcal{X}$，也应指定相应的事件类 (集合类)$\mathcal{A}$ 。在几乎所有的例子中，$\mathcal{X}$ 都是欧氏空间，而 $\mathcal{A}$ 都是一个大的集合类，即 Borel 集合类，它特别包括了所有的开集和闭集。读者若不涉及有关 $\mathcal{A}$ 的结构也不会影响对统计方面内容的理解。

这本书的前身是 1950 年以油印形式出现的讲课笔记，它是由 Colin Blyth 在我于 Borkeley 讲课期间记录的；以这个油印本为基础的讲义随后一直作为这门课的教科书直到本书出版。本书的一些章节是后来由 Michael Stuart 和 Fritz Scholz 修改而成的。在把这些材料写成为一本书的整个过程中，我极大地得益于我妻子 Juliet Shaffer 的支持和建议。Rudy Beran, Peter Bickel,

David Birkes, Colin Blyth, Larry Brown, Fritz Scholz 和 Geoff Watson 阅读了部分手稿，并提出了许多改进意见。6.7 节和 6.8 节分别是基于 Peter Bickel 和 Chuck Stone 提供的材料写成的。特别应该感谢的是 Wei-Yin Loh，他在各个不同的阶段仔细地阅读了全部手稿并检查了所有的习题。他的工作纠正了大量的错误并导致了许多其它的改进。最后，我愿意感谢 Ruth Suzuki 所作的熟练而准确的打字，也愿意感谢 Sheila Gerber 最后的补充和修改。

E. L. Lehmann
Berkeley, California

1983 年 3 月

# 目 录

# 表格目录

# 插图目录

# 第一章　准备知识

## §1.1　问题

统计学是研究数据的搜集、分析和解释的一门学科。本书中，我们将不考虑数据的搜集问题，而是考虑从已给的数据中能够得到什么信息？我们的结论不仅依赖于已得到的数据，而且依赖于得到这些数据的背景知识。后者形成赖以进行分析的假定。根据数据结构的不同假设，有三种不同的处理方法，现介绍如下。

**数据分析**. 基本上不需要对数据作外来的假定，而对数据本身进行分析。数据分析的主要目的是通过综合和整理数据，找出他们的主要特点和弄清楚它们的基本结构。

**经典的统计推断和判决理论**. 经典统计的要点在于假定数据是某些随机变量的观察值，而这些随机变量服从联合分布 $P$, 其中 $P$ 属于某一已知类 $\mathcal{P}$。通常 $\mathcal{P}$ 中的分布用参数 $\theta$ 来标识，其中 $\theta$ 不必为实数，而是取值于集合 $\Omega$, 即

$$\mathcal{P} = \{P_\theta, \theta \in \Omega\}. \tag{1.1}$$

分析的目的就是给出参数 $\theta$ 的合理值 (这是点估计问题), 或者至少决定 $\Omega$ 的一个子集，使得我们有理由断言它包含 (或者不包含) 参数 $\theta$ 的真值 (置信区间的估计或假设检验)。关于 $\theta$ 的这样表述可以视为数据所提供的信息的概括，以供决策者参考。

**Bayes 分析**. 在 Bayes 方法中，尚需补充假定 $\theta$ 自身是一个具有已知先验分布的随机变量 (虽然不可观察)。然后根据观察数据对先验分布进行修正而得到一个后验分布 (给定数据时 $\theta$ 的条件分布), 这个后验分布乃是根据所作的假定以及数据综合出来的关于参数 $\theta$ 的信息。

可以看出，这三种方法所得的结论一个比一个强。这是由于它们是在不同程度的假定之下得到的结论。假定愈强，相应的结论愈具体，但是这种结论的可靠程度就相应减弱。通常，令人满意的做法就是，综合地运用不同的手法来处理同一个问题。例如，我们可以在一组详尽的假定之下设计一项研究 (如决定样本的大小), 而在另一组较弱却显得更可靠的假定之下进行分析。实际上，我们也常常使用几种不同的方法来对同一个问题建立模型。如果的结论具有合理的一致性，则我们就认为所得的结论是令人满意的，反之，则需要对不同的假定进行仔细考察。

本书的第二、三和五章主要采用第二种处理方法，而第四章采用第三种方法。第六章讨论大样本理论，在处理大样本问题时，上述两种方法都涉及到。(关于第一种处理方法的论述，可以参阅 Tukey 的经典著作 " Exploratory Data Analysis "，或者更近代的，由 Hoaglin, Mosteller & Tukey (1985) 写的书，该书包含了 Diaconis 的很有趣的处理方法。) 在本书中，对于给定的问题，我们首先试图说清楚何谓优良的统计方法，然后找出办法以确定优良的统计方法，即求出最优解。

但是这种方法遇到困难。首先，不存在一个公认的优良性准则的定义，其次对于同一个准则，也有不同的认识。例如，即使平方误差损失是一个公认的准则， Bayes 学派，频率学派或条件论学派，对于该准则的认识也必须趋于一致。

更为严重的问题是，最优解和它的性质在很大程度上依赖于所假定的概率模型 (1.1)，而关于模型 (1.1) 的正确性的假定往往是建立在相当脆弱的基础之上的。因此，考虑提出的解在偏离原模型假定的条件之下的 *稳健性* (*robustness*) 就成为重要的问题。一些关于稳健性 (包括 Bayes 观点和频率学派的观点) 的内容将在第四章和第五章讨论。

直到现在为止，进行的是一般性的讨论。现在让我们专门考虑点估计。在模型 (1.1) 中，假定 $g(\theta)$ 是定义在 $\Omega$ 上的实值函数，我们的目的是想知道 $g(\theta)$ (当然 $g(\theta)$ 也可以是 $\theta$ 本身) 的值。不幸的是， $\theta$ (因而 $g(\theta)$) 是未知的。然而我们可以根据数据得到 $g(\theta)$ 的一个估计，希望此估计与 $g(\theta)$ 接近。

点估计是统计推断最常见的形式之一。人们测量一个物理量以估计它的值；进行普查以估计支持某一个候选人的选民人数百分比或者观看某一个电视节目的人数比例；进行农业试验以估计某种新化肥的效用；或者进行临床试验以估计某种医疗措施对病人预期寿命的改进或疾病的治愈率。作为这样的估计问题的典型，我们考虑一个用测量手段来决定未知量的例子。

**例 1.1 测量问题.** 为了估计某个待测量 $\theta$, 如距离 (或温度)，我们对这个量进行多次测量。设 $n$ 个测量值是 $x_1, \cdots, x_n$, 一个常用的估计就是以其均值

$$\bar{x} = \frac{x_1 + \cdots + x_n}{n}$$

作为 $\theta$ 的估计。求一些观察值的平均以得到更精确的值的想法在今天看来是十分平常的，但很难理解人们并非一直利用它。实际上它似乎直到十七世纪末才开始使用 (参见 Plackett 1958)。但是为什么采取平均值这种形式呢？下面给出平均值的两个性质，当初曾经试图用这些来说明这个方法的合理性。

(i) 待测量的一个吸引人的近似值是这样的 $a$ 值，它使得平方和 $\sum(x_i - a)^2$ 达到最小。 $\theta$ 的这个 *最小二乘估计* (*least square estimator*) 就是 $\bar{x}$, 这可以由恒等式

$$\sum(x_i - a)^2 = \sum(x_i - \bar{x})^2 + n(\bar{x} - a)^2 \tag{1.2}$$

看出。由于右端的第一项不包含 $a$, 而第二项当 $a = \bar{x}$ 时达到极小。 (至于最小二乘估计的历史，见 Eisenhart 1964, Plackett 1972, Harter 1974-1976 和 Stigler 1981 。最小二乘估计将在 3.4 节以更一般的形式加以讨论。)

(ii) 在 (i) 中定义的最小二乘估计是使残差平方和达到极小的值，这里残差是指观察值 $x_i$ 与估计值之差。另一种方法是寻找这样的值 $a$, 它使得残差和为零，这样正负残差相互抵消。关于 $a$ 的条件是

$$\sum(x_i - a) = 0, \tag{1.3}$$

这又一次直接得到 $a=\bar{x}$。(当然这两个条件是等价的，其原因是显然的，这是因为 (1.3) 式恰好表示 (1.2) 式关于 $a$ 的导数为零。)

这两个原则显然属于在本节开始提到的数据分析这个范畴。根据这两个原则，可以求出平均值是这组数据的中心位置的一个合理的描述性度量。但是我们并未作出关于 $x_i$ 和 $\theta$ 之间的联系的明确假定，所以我们并不能证明 $\bar{x}$ 作为真值 $\theta$ 的一个估计的合理性。为了建立这样的联系，现在我们假定 $x_i$ 是 $n$ 个独立的，具有依赖于 $\theta$ 的共同分布的随机变量的观察值。Eisenhart (1964) 认为 Simpson (1755) 在为此目的引入概率模型方面迈出了关键性的一步。

特别，我们将假定 $X_i=\theta+U_i$，这里测量误差 $U_i$ 的分布 $F$ 是关于 0 对称的，从而 $X_i$ 的分布关于 $\theta$ 对称的，且具有下列形式

$$P(X_i\le x)=F(x-\theta). \tag{1.4}$$

就这个模型而言，我们现在能否有理由说明平均值比一个独立的观察值提供了更精确的估计值呢？下面是经典统计推断的分析方法。

若 $X_i$ 具有有限方差 $\sigma^2$，则平均值 $\overline{X}$ 的方差为 $\sigma^2/n$，因此 $\overline{X}$ 与 $\theta$ 的差的平方的期望值仅仅是利用一个单独观察值时的 $1/n$。然而，若 $X_i$ 具有 Cauchy 分布，则 $\overline{X}$ 与单独的 $X_i$ 具有相同的分布 (习题 1.6)，因此取一些值加以平均并未多得到什么。这样看来，$\overline{X}$ 是否是 $\theta$ 的一个合理的估计依赖于测量值 $X_i$ 的性质。 □

如上例所示，一个估计问题包括两个部分。

(a) 一个定义于参数空间 $\Omega$ 的实值函数 $g$，其在 $\theta$ 点的值是待估的，我们将称 $g(\theta)$ 为 *待估量* (*estimand*)。[在例 1.1 中，$g(\theta)=\theta$。]

(b) 一个取值于样本空间 $\mathcal{X}$ 的可观察的随机变量 $X$ (通常取向量值)，它的分布 $P_\theta$ 属于 (1.1) 所述的族 $\mathcal{P}$。[在例 1.1 中，$X=(X_1,\cdots,X_n)$，这里 $X_i$ 是独立的同分布的 (iid)，其分布为 (1.4)。$X$ 的观察值 $x$ 构成了 *数据* (*data*)。]

在估计问题中，最关键的是确定一个合适的估计。

**定义 1.2** 在样本空间上定义的实值函数 $\delta$ 称为 *估计* (*estimator*)，它用来估计待估量 $g(\theta)$。待估量 $g(\theta)$ 是参数 $\theta$ 的实值函数。

人们当然希望估计 $\delta(X)$ 与未知的 $g(\theta)$ 相接近，但在估计的形式定义中并不考虑这种要求。当然，实用的估计总是比较合理的估计，因此 $\delta(X)$ 是未知值 $g(\theta)$ 的合理估计。$\delta(X)$ 在 $X=x$ 时的值称为 $g(\theta)$ 的 *估计值* (*estimate*)。

在给出估计的定义时，本来应该采用比定义 1.2 更严格的定义。在应用中，对 $\delta(X)$ 的取值作适当的限制是合乎情理的。例如，当 $g(\theta)$ 仅取正值时，限制 $\delta$ 取正值，当 $g$ 仅取整数时，限制 $\delta$ 取整数值，等等。然而为了数学上的方便，我们暂时不附加这样的限制。

由于 $\delta(X)$ 是一个随机变量，我们只能在某种平均意义上来讨论 $\delta$ 与 $g$ 的接近程度。为使这一论点精确化，必须指定一个估计 $\delta$ 和 $g(\theta)$ 的平均接近的度量 (或者距离)。这些度量的例

子如下：

$$P(|\delta(X)-g(\theta)|<c), \quad \text{对某一个 } c>0, \tag{1.5}$$

和

$$E|\delta(X)-g(\theta)|^p, \quad \text{对某一个 } p>0. \tag{1.6}$$

(其中，我们希望前一项大，而后一项小。) 若 $g$ 和 $\delta$ 都取正值，人们或许感兴趣于

$$E\left|\frac{\delta(X)}{g(\theta)}-1\right|^p.$$

它启发我们将 (1.6) 式推广为

$$k(\theta)E|\delta(X)-g(\theta)|^p. \tag{1.7}$$

更一般地，假定我们用 $L(\theta,d)$ 来度量用 $d$ 估计 $g(\theta)$ 所带来的后果 (损失)。作为 损失函数 (*loss function*) $L$，我们将假定

$$L(\theta,d)\geq 0 \quad \text{对所有的 } \theta,d \tag{1.8}$$

和

$$L(\theta,g(\theta))=0 \quad \text{对所有的 } \theta. \tag{1.9}$$

于是估计到正确值时，损失是 0。一个估计 $\delta$ 的精确程度，更合理地说，不精确程度可以用 风险函数 (*risk functions*)

$$R(\theta,\delta)=E_\theta\{L(\theta,\delta(X))\} \tag{1.10}$$

来度量，它刻划了使用 $\delta$ 作为 $g(\theta)$ 的估计所导致的平均损失。人们希望找到一个估计 $\delta$，使得它对所有的 $\theta$ 值，其风险达到极小。

根据所述的条件，这个问题无解。因为由 (1.9) 式可见，可取 $\delta(x)=g(\theta_0)$ 对一切 $x$，而使得风险在任一给定点 $\theta_0$ 减少到 0。这样就不存在 一致最好的估计 (*uniformly best estimator*)，即不存在这样的估计，使得风险对所有的 $\theta$ 的值同时达到极小，但需排除在 $g(\theta)$ 是常数的不足道的情况。

克服这个困难的一个方法是排除那些有偏的估计，即过于照顾 $\theta$ 的一个或多个值而忽略其他值的那些估计。为此我们要求估计满足某种对参数 $\theta$ 无偏向 (或者公正性) 的条件。一个这样的条件是要求估计的 偏倚 (*bias*): $E_\theta[\delta(X)]-g(\theta)$ 为零 (偏倚亦称为系统误差)，即

$$E_\theta[\delta(X)]=g(\theta) \quad \text{对所有的 } \theta\in\Omega \text{ 成立.} \tag{1.11}$$

总的来看，这个 无偏性 (*unbiasedness*) 条件保证了估计的过高部分和不足部分的平衡。因而估计值在“平均”意义上是正确的。考虑过高与不足估计的频数之间的平衡可以得到另一个无偏性条件，

$$P_\theta[\delta(X)<g(\theta)]=P_\theta[\delta(X)>g(\theta)] \quad \text{对所有的 } \theta\in\Omega \text{ 成立.} \tag{1.12}$$