

现代外国统计学优秀著作译丛

离散多元分析 理论与实践

DISCRETE MULTIVARIATE ANALYSIS
THEORY AND PRACTICE

[美] Yvonne M. M. Bishop

Stephen E. Fienberg 著

Paul W. Holland

张尧庭 译

史宁中 校

中国统计出版社

现代外国统计学优秀著作译丛

离散多元分析:理论与实践

〔美〕 Yvonne M. M. Bishop
Stephen E. Fienberg 著
Paul W. Holland

张尧庭 译
史宁中 校

中国统计出版社

(京) 新登字 041 号

图书在版编目 (CIP) 数据

离散多元分析：理论与实践 / (美) Y. M. M. 比少普
(Bishop, Y. M. M.) 等著；张尧庭译.

—北京：中国统计出版社，1998. 3

(现代外国统计学优秀著作译丛)

书名原文：Discrete Multivariate Analysis: Theory and Practice

ISBN 7-5037-2348-3

I. 离…

II. ①比… ②张…

III. 统计模型：离散模型—多元分析

IV. O212.4

中国版本图书馆 CIP 数据核字 (96) 第 23028 号

著作权合同登记：图字 01-97-0389 号

中国统计出版社出版

(北京三里河月坛南街 75 号 100826)

新华书店经销

科伦克三莱印务(北京)有限公司印刷

*

850×1168 毫米 32 开本 2 2.375 印张 57 万字

1998 年 3 月第 1 版 1998 年 3 月北京第 1 次印刷

印数：1 4 000 册

*

定价：49.40 元

(版权所有 不得翻印)

现代外国统计学优秀著作译丛 专家委员会

主任：

翟立功 国家统计局副局长

副主任：

贺 铿 国家统计局副局长

王吉利 国家统计局统计教育中心主任

委员：

刁锦寰 美国芝加哥大学商学院 教授

吴建福 美国密西根大学统计系 教授

孟晓犁 美国芝加哥大学统计系 博士

张尧庭 上海财经大学数量经济研究所 教授

茆诗松 华东师范大学数理统计系 教授

陈家鼎 北京大学概率统计系 教授

郑祖康 复旦大学统计与运筹系 教授

吴喜之 南开大学数学系 教授

袁 卫 中国人民大学统计系 教授

邱 东 东北财经大学统计系 教授

郝国印 国家统计局统计教育中心副主任

谢鸿光 中国统计出版社副总编

办公室：

刘启荣 国家统计局统计教育中心教材处处长

严建辉 中国统计出版社第二书籍编辑部主任

李 毅 国家统计局统计教育中心教材处副处长

出版说明

为了加强对国外统计理论与实践的研究和了解，全面反映国外统计科研和教学的发展，促进我国统计教学改革和教材内容更新，在国家统计局领导的大力支持下，全国统计教材编审委员会组织翻译出版了这套“现代外国统计学优秀著作译丛”。

随着我国社会主义市场经济体系的逐步建立，统计教育正面临着十分严峻的挑战。一方面，在社会主义市场经济条件下，不论国家的宏观经济调控还是企业的生产经营管理，都要求准确地把握市场运行的态势，科学地分析经济中各种错综复杂的关系，因而，对统计信息的需求越来越大，对统计人才的业务素质提出了更高的要求；另一方面，我国过去的统计教育模式是按为高度集中的计划经济管理体制服务的要求建立的，培养的统计人才的知识结构比较单一，难以适应经济体制、统计体制改革的需要。为使统计人才的培养适应建立社会主义市场经济体制的需要，满足二十一世纪现代化建设的要求，缩小与国际先进水平的差距，基础在教育，关键在教材。在继续组织有关专家、学者编写一批反映国内统计科学和统计实践发展的新教材的同时，必须尽快引进并翻译出版一批外国先进统计教材。这是学习外国先进统计知识的一

种直接而且十分有效的方式，对于推动国内统计教材内容更新和教学改革，造就一大批具有渊博知识和多方面业务技能的复合型人才，具有十分重要的意义。

为了做好这套丛书的翻译出版工作，全国统计教材编审委员会成立了现代外国统计学优秀著作译丛专家委员会，对国外统计著作的出版和使用情况进行了调查研究，分析了国内对外国统计教材的需求，在此基础上制定了翻译著作选题规划。在这套丛书的翻译出版过程中，我们得到了国内外有关专家、有关院校统计系和国外有关出版公司的大力帮助和支持，在此表示衷心的感谢。

全国统计教材编审委员会

1995年7月

译者序

Bishop 等人所著的《离散多元分析》是一本统计学的名著。这本书的特点是对列联表的分析和应用作了全面、详细的介绍，书中例子的丰富、方法的应用、理论的准备，都是各有特色，所以已经多次重印。本书对实际应用和研究人员都有很好的参考价值。

本书实例涉及的面非常广泛，从医院管理、生物遗传，到社会生活、体育运动、政治选举……等等，都有生动而富于启发性的实例。有些实例是对著名的历史资料重作分析，如孟德尔的豌豆试验，军团中被马踢死的士兵数等，都有独到的见解，发人深思。我在翻译的时候，感到十分困难，请教了体育、医药、生物方面的有关专家，但看来还会有不少错译的地方，希望大家看到后给我指出，便于改正。

本书在理论方面，也有专门的几章，放在全书的最后，作了系统的处理。但在不少地方，明确地提出了未解决的问题，希望理论工作者积极参与研究。

我感到本书最大的优点是能对各种方法给以详细的评述，从它的优缺点到适用的范围，应注意的事项，都不厌其烦，多次说明。这对于实际工作者尤有参考价值，一般的统计书籍中，很少有这一类的评述，这充分反映了作者有丰富的实践经验和良好的理论素养。

另一方面，本书的写作往往是夹叙夹议，所以显得头绪繁多，说到一个问题，往往要引前面的，又要说后面什么地方会细谈，所以看起来使人感到常常有小东西塞在中间，打乱了思路，作者是好心，想提醒、照顾读者，这一点是优点，也是可以认为有点不理想。使人感到重复太多，行文较罗唆。我在翻译时就深有此感，

不知是否对，在这里我预先说一下，请读者注意。

书中有一些印刷错误和毛病，我发现了就加以改正，用译者注的方式说明。如果我改错了，注中可以看到原文，以免读者由我误导，我有责任对读者负责。如发现我改错了，欢迎来信指正，将来有机会重印时改过。

我衷心感谢国家统计局教育中心组织翻译一批这样的名著，介绍给广大的统计工作者。我相信此书的出版对推动抽样调查资料的分析会有很好的作用。

感谢中国统计出版社出版这一套丛书。

张尧庭

1996年7月

序 言

自从 Karl Pearson (皮尔逊) 和 Sir R. A. Fisher (费歇) 以来, 多元离散数据的统计分析, 尤其是交叉分类形式的, 在统计文献中占有一个突出的位置。虽然 Maurice Bartlett (巴脱莱特) 关于 $2 \times 2 \times 2$ 表无二阶交互作用的检验的先驱性论文在 1935 年已经发表, 但是多维的交叉分类数据的分析及其方法的使用直到高速计算机的普遍应用之后才有了大幅度的进展。结果是近十年来, 在统计杂志, 如同在生物、社会、医学等杂志一样, 用了很多篇幅来增刊处理多元离散数据分析的论文。许多统计学家对这一进展作出了贡献, 只要对所列的参考文献看上一眼就会立即发现的。特别要指明的, 是 Joseph Berkson, M. W. Birch, I. J. Good, Leo A. Goodman, James E. Grizzle, Marvin Kastenbaum, Gary G. Koch, Solomon Kullback, H. O. Lancaster, Nathan Mantel 和 R. L. Plackett 等人所做的持续而杰出的贡献。

我们最可信赖的、一直在交叉分类数据分析方面做研究的一位学者是 Frederick Mosteller。毫不夸张地说, 没有他的鼓励和多方面对我们的支持, 就不会有这本书。我们对交叉分类数据感到兴趣要追溯到 1964 年, 是 Mosteller 在国家麻醉研究中心工作期间遇到的问题引起的。这些工作直接引出我们之中两位 (Bishop 和 Fienberg) 的博士论文, 以及一系列发表的论文。而 Fred* 对书的贡献远不只是鼓励; 他几乎看过了每一章并且写下了意见, 如

* 指 Mosteller——译者注。

果说有什么优点的话，这都是归功于他的。

Richard Light 热诚地参与了筹划此书的工作，并对若干章节提出了意见。他对第 11 章，关联性和一致性度量准备了初稿，他对这一章的最终格局付出了很大的精力。

我们还要感谢许多我们的同事和学生，他们对手稿的一些部分提出了修改意见，对我们的研究提出了有价值的建议，激励了我们在这一领域的兴趣。我们应感谢的人有 Raj Bahadur, Darrell Bock, Tar Chen, William Cochran, Joel Cohen, Arthur Dempster, O. Dudley Duncan, Hillel Einhorn, Robert Fay, John Gilbert, Anne Goldman, Shelby Haberman, David Hoaglin, Nathan Keyfitz, William Kruskal, Kinley Larntz, Siu-Kai Lee, Lincoln Moses, I. R. Savage, Thomas Schoener, Michael Sutherland, John Tukey, David Wallace, James Warram, Sanford Weisberg, Janet Wittes 和 Jane Worcester 等。

对手稿的付印要感谢 Holly Grano, Kathi Hirst, Garol Lambert 和 Mary Jane Schlepner。

国家自然科学基金对哈佛大学的资助项目 GS-32327X1 对我们研究和写作提供了长期支持。我们还收到了其他科研基金的广泛资助。这些是：国家肿瘤研究所的 CA-06516；国家卫生研究院儿童肿瘤研究基金，资源和研究设备部的 RR-05526；国家科学研究基金 GP-16071, GS-1905，以及从芝加哥大学统计系海军研究室统计组的一个项目，理论生物系 Alfred P. Sloan 基金的一个项目；国家自然科学基金关于国家经济研究所的一个项目和 Paul W. Holland 的社会科学研究院的一个专业研究项目。

一些章节原稿的早期格局在 *The Annals of Statistics*, *Biometrics*, *Biometrika* 和 *The Journal of the American Statistical Association* 等杂志上都刊印过。

Brookline, Massachusetts
New Brighton, Minnesota
Hingham, Massachusetts

Y. M. M. B.
S. E. F.
P. W. H.

1974 年 2 月

目 录

序言

1 引言	1
1.1 需求	1
1.2 为什么写成一本书?	1
1.3 不同的使用者	2
1.4 各章的概述	3
1.5 计算程序	7
1.6 如何使用	7
2 计数数据的结构模型	10
2.1 引言	10
2.2 二维——四格表	13
2.3 二维情形——长方形表	28
2.4 三维表的模型	36
2.5 四维和四维以上的模型	50
2.6 练习	57
2.7 附录: 2×2 表的几何	59
3 完备表的极大似然估计	66
3.1 引言	66
3.2 抽样分布	72
3.3 充分统计量	75
3.4 求极大似然估计的方法	86

3.5	对数线性模型的比例迭代拟合	98
3.6	比例迭代拟合的经典用法	115
3.7	为了拟合模型重排数据	121
3.8	自由度	136
4	正式的拟合优度：综合统计量与模型选择	148
4.1	引言	148
4.2	拟合优度的综合评价	149
4.3	标准化比例	158
4.4	内部拟合优度	164
4.5	选一个模型	189
4.6	附录：Goodman 的分割计算	206
5	不完备表的极大似然估计	214
5.1	引言	214
5.2	不完备的二维表	216
5.3	完备表子集的不完备二维表	249
5.4	不完备的多维表	254
5.5	表示成不完备多维表的二维表	272
6	封闭总体大小的估计	277
6.1	引言	277
6.2	二样的捕获——再捕获问题	280
6.3	N 的条件极大似然估计	286
6.4	三样本调查	288
6.5	一般的多重捕获问题	299
6.6	讨论	310
7	度量变化的模型	312
7.1	引言	312
7.2	一阶马尔可夫模型	317

7.3	高阶马尔可夫模型	325
7.4	有单个转移序列的马尔可夫模型	329
7.5	其他模型	332
8	方表的分析：对称性与边缘均匀性	341
8.1	引言	341
8.2	二维表	342
8.3	三维表	362
8.4	总结	374
9	模型的选取与拟合接近程度的评价：实用方面	375
9.1	引言	375
9.2	建模中的简单性	377
9.3	寻找抽样模型	381
9.4	用同一组数据拟合和检验	384
9.5	拟合过好	391
9.6	大样本以及零假设模型失真时的卡方	397
9.7	不正规数据和参数的缩减	401
9.8	频数分布的频数	406
10	交叉分类中估计和检验的其他方法	416
10.1	引言	416
10.2	信息论方法	418
10.3	极小卡方、修正卡方和对数卡方	423
10.4	逻辑斯蒂模型及如何使用	434
10.5	用分割卡方来检验	438
10.6	基于条件分布检验的精确理论	442
10.7	基于比例变换的分析	446
10.8	必需发展	451
11	关联性和一致性的度量	453

11.1	引言	453
11.2	2×2 表的关联性度量指标	458
11.3	$I \times J$ 表关联性的指标	468
11.4	作为关联性一个特例的一致性	479
12	格子概率的伪贝叶斯估计	489
12.1	引言	489
12.2	贝叶斯估计与伪贝叶斯估计	493
12.3	伪贝叶斯估计的渐近结果	500
12.4	小样本的结论	508
12.5	依赖数据的 λ	512
12.6	另一个例: 二张社会流动表	520
12.7	近期的结论和一些忠告	524
13	离散数据的抽样模型	530
13.1	引言	530
13.2	二项分布	531
13.3	泊松分布	534
13.4	多项分布	537
13.5	超几何分布	545
13.6	多元超几何分布	548
13.7	负二项分布	550
13.8	负多项分布	553
14	渐近方法	556
14.1	引言	556
14.2	O, o 记号	558
14.3	随机序列的收敛性	564
14.4	随机序列的 O_p, o_p 记号	579
14.5	矩收敛性	591
14.6	计算渐近分布的 δ 法	593

14.7 多项估计和检验的一般框架	613
14.8 多项最大似然估计的渐近行为	620
14.9 多项拟合优度检验的渐近分布	626
参考文献	649
数据集索引	676
名词索引	685

1 引言

1.1 需求

科学家对大量数据的系统探索结构时，会用他自己的准则去寻找灵感，从现代的计算技术中得到支持，从统计模型中得到引导。因为大量的数据看来是复杂的，对它们可以用很多的方法来进行分析，汇编这些分析方法，提供给科学家值得去尝试的方向，这是一条有吸引力的捷径而不会是一种束缚。在统计专题中，如回归分析、方差分析都已汇编成广泛使用的，有时是重要的原则。在离散多元统计方面，无论如何，已经有了可用的很好的入门著作，如 Cox [1970]，Fleiss [1973]，Good [1965]，Lancaster [1969] 以及 Maxwell [1961]，但在面对大列联表的系统处理时就突然停住了，特别是面对一些麻烦的不规则的表时更是如此。本书则提供了这样的一类处理。

1.2 为什么写成一本书？

离散多元分析虽然有充实的文献，不幸的是散见在各处。这本书将这些文献汇集成有组织的内容。尽管我们确实在书中有一