



中青年经济学家文库

ZHONGQINGNIAN JINGJIXUEJIA WENKU

教育部人文社会科学研究青年基金项目(09YJC910002)资助成果

多重插补方法与 应用研究

庞新生 / 著

DOCHONG CHABU FANGFA YU YINGYONG YANJIU



经济科学出版社
Economic Science Press

中青年经济学家文库

教育部人文社会科学研究青年基金项目（09YJC910002）资助成果

多重插补方法与 应用研究

庞新生 著

经济科学出版社

图书在版编目 (CIP) 数据

多重插补方法与应用研究/庞新生著. —北京:
经济科学出版社, 2013. 6

ISBN 978 - 7 - 5141 - 3841 - 2

I. ①多… II. ①庞… III. ①经济统计 -
统计方法 - 研究 IV. ①F222. 1

中国版本图书馆 CIP 数据核字 (2013) 第 233321 号

责任编辑: 刘怡斐

责任校对: 隗立娜

责任印制: 邱 天



多重插补方法与应用研究

庞新生 著

经济科学出版社出版、发行 新华书店经销

社址: 北京市海淀区阜成路甲 28 号 邮编: 100142

总编部电话: 010 - 88191217 发行部电话: 010 - 88191522

网址: [www. esp. com. cn](http://www.esp.com.cn)

电子邮件: [esp@ esp. com. cn](mailto:esp@esp.com.cn)

天猫网店: 经济科学出版社旗舰店

网址: [http: //jjkxcbs. tmall. com](http://jjkxcbs.tmall.com)

北京万友印刷有限公司印装

880 × 1230 32 开 5.25 印张 200000 字

2013 年 9 月第 1 版 2013 年 9 月第 1 次印刷

ISBN 978 - 7 - 5141 - 3841 - 2 定价: 20.00 元

(图书出现印装问题, 本社负责调换。电话: 010 - 88191502)

(版权所有 翻印必究)

前 言

缺失数据的处理是抽样推断中无法回避的难题，作为处理缺失数据非常有效的方法，多重插补方法的问世具有重要的意义，该方法在国内的实践应用尚处于初始阶段，仍有很多问题值得进一步探讨。本书在借鉴国内外相关科研成果的基础上，结合国内调查实例，讨论多重插补在抽样调查中缺失数据处理的实现过程。缺失数据涉及范围很广泛，从来源看，既包括实验中的缺失数据，也包括调查中的缺失数据。由于实验受人为因素影响很小，其中的缺失数据处理相对要简单一些，而调查受主观因素影响很大，处理比较复杂。因此，本书主要着眼于调查中出现的缺失数据，对缺失数据多重插补方法理论和应用进行了系统回顾和研究，并对某些多重插补方法仅作了尝试改进，希望本书对于国内缺失数据处理方法的后续研究有一定的参考价值。

本书得的出版来自于教育部人文社会科学研究青年基会项目(09YJC910002)的鼎力资助。尽管在此之前笔者也进行过一些有关缺失数据多重插补方法相关问题的研究，但这些研究仅局限于2004年以前国内外相关成果，且没有形成体系，项目的设立为本研究完成后续研究创造了条件。本书写作过程所涉及的诸多内容得益于恩师金勇进教授的悉心指导，值此书出版之际，谨向恩师致以深深地敬意和由衷的谢意。感谢中国人民大学的蒋妍博士无偿提供了大量的数据，感谢厦门大学的王华博士对于项目研究的大力支持与参与。衷心感谢北京林业大学的陈建成教授在出版资会上

的支持和资助，使得本书得以顺利出版，感谢所有曾经帮助和鼓励过我的老师与朋友，项目得以顺利完成，离不开大家的帮助和支持。特别值得一提的是经济科学出版社刘怡斐副编审给予的帮助，没有她和出版社工作人员的辛勤的工作，这本书就难以面世。感谢美国宾西法尼亚州立人学的 Joseph L. Schafer, Ph. D, 书中所使用的关于多重插补的部分软件是从他的个人网页下载的。笔者在写作过程中引用了许多参考文献中的例子，在此向有关作者表示感谢。

诚然，限于本人的水平，书中难免挂一漏万，尚存诸多不足之处，恳请各位前辈学者和年轻学友斧正。

目 录

第 1 章 导论	1
1.1 概述	1
1.2 多重插补研究综述	2
1.3 本书的结构	10
第 2 章 多重插补方法的基本问题	13
2.1 缺失数据的基本问题	13
2.2 多重插补方法及其统计推断原理	23
2.3 单调缺失数据的多重插补	43
第 3 章 复杂情况下多重插补方法	46
3.1 不可忽略机制下缺失数据的多重插补方法	46
3.2 考虑抽样机制条件下缺失数据的多重插补方法	51
3.3 任意缺失模式下的多重插补方法	55
3.4 多重插补方法的算法实现	58
第 4 章 一种常用多重插补方法的改进	67
4.1 基于多变量事后分层基础上的多重插补方法的基本思想	67
4.2 基于多变量事后分层基础上多重插补方法的算例	75
4.3 基于多变量事后分层基础上多重插补方法的特点	88

第 5 章 多重插补方法效率研究 91

 5.1 缺失数据处理方法的比较 92

 5.2 多重插补方法的效率研究 106

 5.3 采用相关方法处理缺失数据的一个例子 110

 5.4 采用插补方法处理缺失数据的实例比较 119

第 6 章 多重插补方法实证分析..... 128

 6.1 多重插补统计软件的比较分析 128

 6.2 实际调查数据的多重插补处理实证分析 144

主要参考文献..... 154

第 1 章

导 论

1. 1

概述

数据缺失是抽样调查中经常出现的现象，它的出现造成一系列问题：效率降低、数据处理和分析变得复杂、由于观察值与未观察值间存在差异而导致估计出现偏倚。对于统计分析人员来说，如何选择合适的方法对不完整的数据集进行分析，始终是数据处理的难题，关于缺失数据统计处理方法的研究一直是统计学研究中一个活跃领域，并且出现了很多行之有效的处理方法，20 世纪 70 年代出现了热卡法（ad hoc procedures）、删除不完全单位（case deletion）、单一插补（single imputation）；20 世纪 80 年代出现了似然估计（like-based estimation procedures）、EM 算法；20 世纪 90 年代出现了多重插补（multiple imputation, MI）、MCMC 法（Markov Chain Monte Carlo）及贝叶斯方法（Bayesian methods）。在多重插补出现以前，列表删除（list-wise deletion，即删除含有缺失值的所有个案）、成对删除（pair-wise deletion，即只删除分析变量内的缺失值）和单一插补（single imputation）是处理缺失值问题的主要方法，但是这些方法都没有考虑到缺失数据的不确定性以及缺失数据与观察到的数据间可能存在的系统性差异，所以难以提供关于总体参数的准确估计，而多重插补很好地弥补了它们的缺陷，即使插补

模型选择不合理,也能获得比较准确的估计结果。多重插补法的出现,从根本上改变了统计学家对缺失数据的看法,从此以后,缺失数据由简单的剔除或用单一插补进行处理,上升为一种“平均”的变异性,该方法建立在统计原则的基础上,充分考虑了由于缺失数据带来的推断的不确定性,认为回答数据可能隐含着缺失数据的概率分布,一些人甚至把多重插补看成是统计学的魔术,好像可以使信息从无变有,事实上,多重插补只是一种表现缺失数据不确定性的工具。多重插补方法出现之初,由于计算过程极为复杂,因此,在当时并没有得到广泛的应用,在 20 世纪 80 年代,插补缺失数据的方法只是局限于极大似然估计和 EM 算法,直到 20 世纪 90 年代,随着新的计算方法和统计软件的出现,多重插补才成为处理缺失数据的主要工具。目前,国内相关研究比较少,随着市场经济的发展,在不断密切的国际往来过程和宏观微观管理的科学化、现代化进程中,国内各式各样的抽样调查和普查工作成百倍上千倍不断增加,传统缺失数据的处理方法已经无法满足政府统计和商业调查的需要,如何能够保证估计结果的准确性和合理性,显得格外重要,因此,对于多重插补方法的深入研究和推广使用具有相当的现实意义。

1.2

多重插补研究综述

20 世纪 70 年代末, Dempster, Laird 和 Rubin (1977) 提出一种有效估计不完全数据算法——EM 算法,基于这一算法, Donald B. Rubin (1977 年) 提出了通过模拟的思想来实现缺失数据的多重插补,此后, Rubin 发表一系列论文,并于 1987 年将其整理出版,系统的归纳了多重插补方法思想和理论框架。从此,陆续出现关于多重插补方法的进一步的理论发展与应用研究,据不完全统计大约有近五百多篇。这方面的经典著作主要有: Dempster, Laird 和 Rubin (1977) 关于 EM 算法的论文, Little 和 Rubin (1987) 关于

缺失数据的著作, Rubin (1987) 关于多重插补的著作, Schafer (1997) 关于多元缺失数据分析的著作。

1.2.1 关于多重插补方法的理论研究

关于多重插补理论研究主要集中在以下几方面:

第一, 关于多重插补基本问题的讨论, 如多重插补的界定, 多重插补的建立和结果合并, 多重插补次数, 插补模型, 多重插补与单一插补、EM 算法、MCMC 算法的比较等。Rubin (1978) 从贝叶斯统计的角度提出了多重插补基本思想, 对作为结果的多重插补数据集的分析被看做是插补模型下预测分布的模拟, 将多重插补过程归结为插补任务、估计任务和建模任务, 并指出前两项任务容易实现, 可忽略无回答机制下, 建模比较容易, 而不可忽略无回答机制下, 非常复杂。Herzog 和 Rubin (1983), Rubin 和 Schenker (1986), Raghunathan 和 Rubin (1987) 指出, 由于多重插补考虑了区间估计和显著性水平的变异性, 即使对每个缺失值仅插补几次都要优于单一插补, Rubin (1988) 就多重插补与单一插补进行详细的比较; Li, Raghunathan 和 Rubin (1991) 探讨了多元缺失数据的点估计和协方差的合并问题; Li 等 (1991) 探讨了多元缺失数据的 p 值合并问题; Meng 和 Rubin (1992) 探讨了多元缺失数据的似然率检验统计 (likelihood-ratio test statistics) 的合并问题, Meng (1995), Rubin (1996) 和 Schafer (1997) 就插补模型与分析模型的关系问题做了大量的讨论。1996 年 Rubin 在美国统计协会上发表论文 *Multiple imputation after 18 + years*, 澄清了人们对多重插补法的一些错误的认识, 并将多重插补法与其他一些处理数据的手段如加权法等, 做了比较, 论证多重插补法的优越性。Mingxiu Hu, Sameena M. Salvucci, Michael P. Cohen (1998) 对一些常用的插补算法进行了评价。Maren K. Olsen (1998) 在 *Multiple Imputation for Multivariate Missing—data Problems* 中提出了多变量缺失值的多重插补方法,

并进行了实际应用。Gabriele B. Durrant (2006) 针对政府调查中的无回答现象造成的数据缺失做了讨论并强调了在这一领域的一些新发展, 在一个案例中对插补和加权这两种方法做了比较, 同时也说明了缺失数据与测量误差间的密切关系。Susanne Rassler 和 Regina T. Riphahn (2006) 也对这两种方法进行了比较分析, 不仅描述了基本方法和各自的优缺点同时用模拟研究说明了其效应。Ulrich Rendtel (2006) 梳理了有关缺失数据和多重插补问题的已有研究内容, 为后续研究提供了方向。Donald B. Rubin (2006) 对其中有关创造性的统计工作应致力的方向做了进一步的研究, 介绍了基于多重插补的有效区间估计与评估假设的相关内容。

第二, 关于大样本、小样本模式下多重插补的讨论。Graham, J. W. 和 Schafer, J. L. (1997) 将多重插补法的应用范围扩大至小样本; Barnard, J. and Rubin (1999) 给出关于小样本多重插补估计量合并原则, 调整了小样本下 t 分布自由度。

第三, 关于多重插补参数估计和估计方差的讨论。关于插补方差估计方法有三种: 第一种是多重插补 (Rubin, 1977, 1987); 第二种是通过模型辅助的方法 (Särndal, 1990); 第三种是刀切法 (Rao, 1992)。Lee, Rancourt 和 Särndal (1994) 讨论了采用最相邻插补 (the nearest neighbour imputation) 进行多重插补方差估计问题, 指出由于很难给出最相邻插补的适当多重插补, 方差估计往往是偏低的, 并采用模型辅助方法构造出更适合最相邻插补的比率插补, 尽管效果优于多重插补, 但偏差仍存在, 且是不可忽略。Kovar 和 Chen (1994) 讨论了刀切法在最相邻插补方差估计中的应用, 由于调整方法不够理想, 所以尽管减小标准完全数据的方差, 但没有彻底消除。Eric Rancourt, Carl Särndal 和 Hyunshik Lee (1994) 采用以模型辅助的方法给出基于线性回归模型的最相邻插补法正确的方差估计。Fay (1991, 1992, 1994) 通过简单事例表明在某些数据集 (如 mass imputation) 采用多重插补直接估计的方差估计并非一致。J. M. Robins 和 N. Wang (2000) 提出了关于单一

插补和多重插补的近似方差新的估计。Rao (1992) 提供了基于调整插补值的单一插补刀切法方差估计以及在随机缺失机制下近似一致的线性化刀切法方差估计。

第四, 关于常用抽样模式下多重插补的讨论, 如对分层随机抽样、多阶段抽样、不等概抽样情况下多重插补的讨论。Eugene M. Burns (1989) 将多重插补应用于能源信息部的商业建设能源消费调查 (CBECS) (1988, 1989) ——多阶段地区概率抽样, 采用刀切法进行参数估计和估计方差的计算。Burns (1990) 使用准重复热层插补为多阶段调查提供刀切法方差估计, Rao 和 Shao (1992) 指出 Burns 提出的刀切法方差估计可能会造成严重的高估。Binder, D. A. 和 Sun, W. (1996) 则专门研究了复杂样本设计中多重插补的频数推论问题。

第五, 关于常用插补模型的讨论, 如线性回归插补模型下多重插补的实现。Laura C. Lazzeroni, Nathaniel Schenker 和 Jeremy M. G. Taylor (1990) 研究了模型设定错误情况下, 基于正态线性回归模型, 以模型为基础的方法 (model-based method)、残差抽取方法 (residual-draw method)、预测均值匹配法 (predictive-mean-matching method)、局部残差抽取法 (local residual draw method) 四种多重插补的稳健性问题。Brownstone D. (1991) 介绍了线性回归插补模型下多重插补的实现。Robert Zeithammer 和 Peter Lenk (2006) 对缺失维度的多元正态模型做了相关研究, 由于缺失维度会使得标准的贝叶斯分析的多维协方差结构变得复杂, 对此提出了一个便于分析和估计的方案, 为我们了解多重插补的相关理论提供了基础。

第六, 对不可忽略缺失机制下多重插补方法的讨论。在研究血压数据时, Cook, N. R. (1997) 提出了不可忽略缺失数据机制下处理多重插补的一种方法。

第七, 关于特殊数据类型下多重插补的实现以及一些特殊方法在多重插补中的应用, 如蒙特卡洛马尔科夫链、半参数法、刀切法的应用。Lipsitz, N. R., Zhao, L. P. 和 Molenberghs, G. (1998)

与 Schenker, N. 和 Tayler, J. M. G. (1996) 将半参数方法引入多重插补, 解决其复杂的建模问题。Schafer 和 Olsen (1998) 采用 DA 算法 (data augmentation) 实施多重插补, Schafer (1997) 论证了采用 DA 算法只需要很少的插补次数就能够获得理想的结果。Landrum MB 和 Becker MP (2001) 探讨了多重插补在不完全纵向数据中的实现。Michael P. Cohen (1997) 对各种类型的贝叶斯自助法 (包括 efron's bootstrap, rubin's Bayesian bootstrap, the finite population bootstrap of Gross, and the finite population bayesian bootstrap of loare) 作了简单总结, 运用不等概率贝叶斯自助法进行多重插补。Schafer 和 Olsen (1999) 讨论了半连续变量的模型和插补问题。Daniel Schunk (2008) 在相关著作中介绍了蒙特卡罗的马氏链的理论基础, 并概述对社会经济数据集运用 MCMC 算法的重要技术。Ting Hsiang Lin (2010) 对 EM 算法和 MCMC 算法做了比较, 在数据缺失的两种极端情况下这两种算法没有显著区别。

第八, 一些批评的观点。Robert E. Fay (1992) 指出多重插补在实际应用中的关键局限性, 而且讨论当多重插补失效情况下提供合理推断的备选方法。Phillip S. Kott (1995) 就“适当的多重插补估计是无偏的”这一观点提出一些相反的看法。

1.2.2 关于多重插补方法的应用研究

关于多重插补法的应用文献也有不少。比较有影响的应用主要有: Thomas N. Herzog 和 Rubin 将多重插补用于美国人口普查局 (CPS) 的社会安全福利数据, 验证多重插补法的功效; Daniel F. Heitjan 和 Roderick J. A. Little (1989) 将多重插补用于致命事故报告系统中缺失数据的处理; Clogg, C., D. Rubin, N. Schenker 等 (1991) 将多重插补法用于插补工业和职业普查中的缺失数据; Steven G. Heeringa (1994) 将多重插补法用于插补健康和退休调查中的项目缺失数据; EzzatiRice, T. M., Johnson, W., Khare, M.,

Little, R. J. A., Rubin 和 Schafer (1995) 将多重插补用于健康调查数据; Barnard, J. 和 Meng, X. L. (1999) 则在 Aids 与 NHANES 病理的研究中使用了多重插补技术; Zhou XH, Eckert GJ 和 Tierney WM. (2001) 在公共健康研究中使用了多重插补; Niels Smits, Don Mellenbergh 和 Harrie Vorst (2002) 在学校成绩记录研究中采用多重插补方法处理无法获得的成绩, 并且对 EM 算法、DA 算法作了比较; Fiona M Shrive, Heather Stuart, Hude Quan, William A Ghali (2006) 比较了运用六种不同的估算方法处理曾氏自评抑郁量表 (SDS) 中的缺失数据所得的结果。这六种估算方法分别是: 多重插补、一元回归、个别平均值、整体平均、参与者以前的回答和从 1 到 4 中随机选择的值。结果显示多重插补是较为准确的方法。Miao Jiang, E. Michael Foster 和 Christina M. Gibson-Davis (2010) 用倾向得分法分析了母乳喂养的开始和持续时间对儿童发展的影响。Nicholas T. Longford, Catia Nicodemo, Montserrat Nu'ñez 和 Esther Nu'ñez (2011) 将倾向得分与无回答的多重插补相匹配来评估类风湿关节炎患者的幸福感与肥胖的关系。Jörg Drechsler (2011) 对德国就业研究所的一个实际案例做了研究, 不仅讨论了可行的插补方法, 还对插补软件所能解决的问题做了总结。Chang-Keun Han 和 Song-lee Hong (2011) 研究了韩国 50 岁以上的老年人生活满意度与资产情况的关系, 对这一问题的面板数据进行多重插补后建立了多元回归模型。Yin Bun Cheung, Rhian Daniel 和 Gim Yew Ng (2006) 在新加坡这个拥有保守的性态度的多种族社会做了有关对性生活满意度的问卷调查, 并利用平均插补法和多重插补法处理该问题, 比较发现结果类似。在流行病学调查中, 缺失数据的问题为越来越多的人所重视, 基于这方面案例的多重插补研究也越来越多, Yulei He, Recai Yucel 和 Alan M. Zaslavsky (2008) 基于癌症登记数据库对多重插补方法做了相关介绍; H. Toutenburg, V. K. Srivastava, Shalabh (2008) 介绍了利用插补法对不完整数据的均值的估计; A. M. Brewster, T. J. Jorgensen (2006) 等指出利用

插补方法处理缺失数据得到的结果优于省略缺失数据；Michael R. Elliott（2008）在关于儿童肥胖研究中运用多重插补描述录入误差。Federica Barzi（2006）等提出多重插补相较于简单的插补方法是更有效的估计方法，它在保留所观察到的数据的分布和变量关系的同时进行了合理估计。Susan M. Shortreed, Eric Laber, Daniel J. Lizotte 和 T. Scott Stroup（2010）等在讨论强化学习在优化慢性疾病治疗政策中发挥的作用时，运用多重插补方法解决数据缺失的问题。Diane L. Fairclough（2008）等用 HIV 临床试验作为案例，描述了基于完全随机缺失（MCAR）、随机缺失（MAR）和非随机缺失（MNAR）三种假设条件下几种不同的处理缺失数据的分析方法，包括 MCMC 多重插补法。Ivo M Foppa 和 Md Monir Hossain（2008）采用分层建模的形式，利用 MCMC 插补法评估流感对公众健康造成的影响。Yongwen Jiang 和 Jana Earl Hesser（2006）利用多重插补的方法处理来自罗德岛危险行为因素监测系统（BRFSS）中的缺失数据，用 10 个与健康相关的生活质量指标建立因素 logistic 回归模型进行分析，以了解人口、健康风险行为以及生活品质之间的因果关系。Shinichi Nakagawa 和 Robert P. Freckleton（2011）用一个行为生态学例子说明缺失数据如何影响模型选择或基于假设检验的结论，展示利用多重插补的方法如何将缺失数据恢复并对目标参数做出准确的估计。Filia Vonta 和 Alex Karagrigoriou（2007）详细讨论了在各种生存模型中变量的选择策略以及它们的作用，指出多重插补方法能够成功的处理审查意见并构造一个竞争对手。Martijn W Heymans, Stef van Buuren, Dirk L Knol, Willem van Mechelen 和 Henrica CW de Vet（2007）讨论了预后研究中的变量选择问题，认为多重插补与引导相结合的方法选择变量，解决了数据集有缺失值的问题同时能够使多变量模型具有良好的性能。John W. Graham, Allison E. Olchowski 和 Tamika D. Gilreath（2007）用蒙特卡罗模拟的方法对插补的次数进行了讨论和说明。Andrea Marshall, Douglas G. Altman 和 Patrick Royston 等（2009）结合多重插补的方法对预后模型

中的利率估计做了总结研究,通过对已有文献的梳理介绍了具体的方法和指导方针。James Alan Fox 和 Marc L. Swatt (2009) 利用多重插补的方法对补充凶杀案报告做了分析,探讨了对受害者、罪犯、和事故特征分别赋权下的影响。由此,可以看出,多重插补法处理缺失数据技术在国外的应用已经是相当广泛,深入到了各行各业。

尽管多重插补在 1977 年就由 Rubin 提出,但由于计算过程极为复杂,并没有在当时就得到广泛的应用。直至 20 世纪 90 年代,随着计算机推广普及,多重插补才成为处理缺失数据的主要工具。这一时期关于多重插补在计算机上的研究包括: Mingxiu Hu, Sameena Salvucci, Stanley Weng (1996) 比较了 Proc Impute 软件 (PC version 2.0, 1991) 与 Schafer 多重插补软件 (version 2.0, 1995)。1997 年 Statistical Solutions Ltd. 出版了 SOLAS 软件,专门用于处理含有缺失值的数据库,也可以实现对缺失数据的多重插补。Schafer (1997) 设计了 NORM、CAT、MIX 和 PAN 等多重插补专业软件,并出版了相关的著作 (*Missing Data in Multivariate Analysis*)。Yang, C., Yuan 和 Rockville, M. D. (2000) 探讨了多重插补在大型统计分析软件 SAS 上的实现。Horton 和 Lipsitz (2001) 进行了针对缺失变量的回归模型软件比较。

1.2.3 国内关于多重插补方法的研究

无论是中国大陆还是中国台湾,关于多重插补方法的研究都比较少。在我国台湾,赵明德及谢邦昌 (1999) 在《探索真像——抽样理论与务实》一书中对多重插补法做了相关介绍,李泰明、郑宇庭和谢邦昌 (1999) 采用 Gibbs 抽样的随机模拟过程讨论多重插补过程中抽样问题,结合蒙特卡罗 EM 算法提出一个一般化的解决方法,并应用实例证明其实用性。另外,还有一些相关主题的学术演讲,主要有:台湾大学数学系陈宏教授 2001 年 4 月在

台湾中央研究院统计科学研究所做的 On Obtaining Correct Variance for Hot-Deck Imputation 学术演讲以及台湾中央研究院统计科学研究所程毅豪博士 2003 年 4 月在台湾卫生研究院所做的 Estimation in Generalized Linear Mixed Models with Multiple Imputation 学术演讲。

中国大陆目前关于抽样调查中缺失数据多重插补处理方法的研究，无论是在理论探讨上还是实际应用方面，远不及国外研究，在极为有限的关于多重插补研究的学术文献中，较有代表性的有：蒋妍（2001）研究了调查中缺失数据的处理方法，其中对多重插补法做了一些介绍；张琅（2002）对多重插补的一些基本机理做了介绍，曹阳、谢万军和张罗漫（2003）对多重插补方法及其统计推断原理做了简要介绍。

综上所述，从国内外关于缺失数据多重插补方法的研究和应用格局来看，至少可以得出以下两点结论：第一，国外关于抽样调查中多重插补方法的基本理论框架已经形成，目前的前沿性研究或者是进行方法的改进或比较研究，或者是利用已有方法进行应用研究和实证分析，还有一个非常显著地特点就是关于多重插补处理软件的研究和比较也是层出不穷，而且一些比较知名的统计软件也在着手完善关于多重插补的模块。第二，台湾的有关研究除了进行引进工作外，基本上遵循了目前国外研究的思路，相比之下，大陆目前相关研究还处于探索阶段，特别需要在概念、方法和应用方面进行介绍和研究。

1.3

本书的结构

由于缺失数据涉及范围很广泛，从来源看，既包括实验中的缺失数据，也包括调查中的缺失数据，由于实验受人为因素影响很小，其中的缺失数据处理相对要简单一些，而调查受人的因素影响