



当代统计方法与应用丛书

李金昌 主编

统计数据质量诊断 与管理研究

TONGJI SHUJU ZHILIANG ZHENDUAN

YU GUANLI YANJIU

程开明 著



浙江工商大学出版社
Zhejiang Gongshang University Press

统计数据质量诊断与管理研究

TONGJI SHUJU ZHILIANG ZHENDUAN YU GUANLI YANJIU

程开明 著

 浙江工商大学出版社

图书在版编目 (CIP) 数据

统计数据质量诊断与管理研究 / 程开明著. —杭州:
浙江工商大学出版社, 2010. 12

ISBN 978-7-81140-260-5

I. ①统… II. ①程… III. ①统计数据—质量管理—
研究 IV. ①C816

中国版本图书馆 CIP 数据核字 (2010) 第 239405 号

统计数据质量诊断与管理研究

程开明 著

责任编辑 白小平 许 静

责任校对 周敏燕

封面设计 刘 韵

责任印刷 汪 俊

出版发行 浙江工商大学出版社

(杭州市教工路 198 号 邮政编码 310012)

(E-mail: zjgsupress@163.com)

(网址: <http://www.zjgsupress.com>)

电话: 0571-88904980, 88831806 (传真)

排 版 杭州中大图文设计有限公司

印 刷 杭州广育多莉印刷有限公司

开 本 880mm×1230mm 1/32

印 张 8.25

字 数 257 千

版 印 次 2010 年 12 月第 1 版 2010 年 12 月第 1 次印刷

书 号 ISBN 978-7-81140-260-5

定 价 25.00 元

版权所有 翻印必究 印装差错 负责调换

浙江工商大学出版社营销部邮购电话 0571-88804227

总 序

统计学的突出特征是通过数据分析得到客观而真实的结论,帮助人们提高对研究对象的认识,因而统计学是探求真理不可缺少的工具。作为定量认识的有力工具,统计方法适用于对自然、社会、经济等广泛现象的探索,具有普适性。正如著名统计学家 C. R. 劳所说:当今,在人类活动努力的一切范围内,统计学已经成为一种万能的、强有力的和不可缺少的研究工具。

随着科学技术的发展,统计学也在不断吸收、借鉴和融合其他学科知识的过程中,深化、丰富和完善其自身理论,不仅应用领域进一步拓展,而且为推出新技术和新方法提供了很好的契机和平台。与此同时,社会经济发展过程中出现的一些新情况、新问题,也对统计学发展提出了新的、更高的要求。当前,就统计学的研究对象而言,所涉及的范围如此之广,所面临的“未知”如此之多,迫使我们不断夯实和拓展统计学科基础,进一步加强统计方法及其应用研究,并把相应的研究成果加以总结和整理。这正是我们依托浙江省高校人文社科重点研究基地(浙江工商大学统计学)编撰“当代统计方法与应用”丛书的目的所在。

浙江工商大学统计学专业是学校传统优势学科之一,在国内相同领域具有较大的影响力。学科于 1980 年开始招收本科生,1990 年获准设立硕士学位点,1996 年被评为浙江省重点扶植学科,2002 年被评为浙江省重点学科和重点专业,2003 年获准设立博士学位点,2004 年被列为浙江省首批省属高校人文社科重点研究基地,2007 年入选国家特色建设专业。迄今为止,浙江工商大学已经为社



会培养了约 1500 名统计学本科生、200 多名统计学硕士生和 20 余名统计学博士生。

学科、专业的建设与发展有赖于团队建设和科学研究。在大家的共同努力下,浙江工商大学统计学科团队成为了省级教学团队和省级科研创新团队,在教学与科学研究方面取得了大量的成果。近五年来,团队在《统计研究》、《经济研究》、《管理世界》、《财贸经济》、《数量经济技术经济研究》等一流杂志及核心期刊公开发表学术论文 150 多篇,先后承担了国家自然科学基金 1 项、国家社会科学基金 7 项,省自然科学基金 3 项、省哲学社会科学规划项目 8 项,其他省部级项目 30 余项,累计项目总经费达 800 多万元,获省部级二、三等奖项 20 多项。目前,浙江工商大学统计学已基本形成具有较强影响力的“统计基本理论与方法”、“经济统计方法与应用”、“管理统计方法与应用”、“概率统计方法与应用”等研究方向。

“当代统计方法与应用”丛书涵盖浙江工商大学统计学科团队成员的科研著作、博士论文,以及基地立项项目的最终研究成果。希望该丛书的出版,有利于推动统计方法的创新,有利于拓宽统计应用的领域,有利于提升统计分析的高度,为促进我国统计学科的建设与发展起到应有的积极作用。

李金昌
二〇一〇年春

前 言

近年来,政府统计部门为提升和改进数据质量付出了艰辛的劳动,进行了不懈的努力,但数据质量仍是困扰我国统计工作的重要问题之一。数据质量的好坏决定着数据分析结果是否客观、真实,进而影响到决策的科学性;数据质量是政府统计能力建设的重要内容,是提升统计公信力的重要源泉。分析显示,多方面的原因导致我国统计数据质量仍受到较多质疑,其中数据质量诊断与管理水平相对落后无疑是一个重要因素。面对这种情况,迫切要求统计学者和实际工作者,多方探求统计数据质量诊断与管理的理论、方法及措施。

立足于数据收集过程,各界已提出许多提高数据质量的途径,但对于如何在现有数据条件下,使数据更能满足统计分析要求,提高分析结果的客观性、可信度,相关的研究还不多。学界普遍重视对数据收集和数据分析的研究,相对忽视数据收集之后、正式分析之前这一过程。基于此,本书上篇立足于数据预处理视角,探讨统计数据质量诊断的方法及应用,希望从中找到提升数据质量的有效策略。

无论是国际组织,还是各国统计机构都十分重视统计数据质量的评估与管理,相继提出了一些统计质量保证原则和基本框架,建立起数据质量评价机制和管理体系。这些统计质量管理框架倡导多维度的数据质量特性、加强统计过程监控与产品评估以及全面质量管理理念,在很多方面表现出较好效果。所以,本书下篇对国际上统计质量管理体系的质量概念、框架、政策和做法加以概括、综合,从中得到加强我国统计数据质量管理的若干借鉴与启示。

由于统计数据质量涉及到数据的生产者、使用者、提供者等直接



参与主体及众多的外部监督主体,并非统计部门一家所能决定;不同主体对数据质量具有不一样的理解,所关注的质量维度也各有特点。所以,开展统计数据质量管理需要综合考虑不同主体的不同需求。本书下篇引入利益相关者理论,全方位地探讨统计数据质量管理框架、评估方法及对策,有利于明确政府统计数据质量管理的方向和重点。

本书的研究内容共包括总论、上篇、下篇及结语四大部分,其中上篇“统计数据质量诊断”包含六章内容,下篇“统计数据质量管理”包含四章内容,全书共计十二章。

第一章为“总论”。立足于有关中国统计数据质量的论争,对国内外统计数据质量的相关研究进行系统综述;接着从统计数据质量的概念着手,整合各方对数据质量维度的理解,并从外部环境和统计系统自身两个方面探析我国统计数据质量问题产生的原因。

第二章为“统计数据质量诊断的新视角”。首先阐明开展统计数据质量诊断的必要性,提出统计数据质量诊断的新视角——数据预处理;然后给出数据预处理的四大过程,对照统计数据预处理过程,提出数据预处理的六大类方法体系。

第三章为“统计数据的初步审查”。概括了用以开展统计数据初步审查的描述及探索性方法,包括频数分析、集中和离散趋势、分布形状特征、列联表、统计图表等分析方法,简述对调查数据进行信度和效度检测的原理,展示了各种方法的实际应用。

第四章为“统计数据的缺失值处理”。以缺失数据的产生机制及表现模式为基础,从“事前”和“事后”两个角度提出缺失数据的处理方法;事前处理包括一些预防性措施和对敏感性问题的脱敏处理方法,事后处理则包括个案剔除法、加权调整法和插补法,不同的方法各有其优缺点。

第五章为“统计数据的异常值处理”。从异常值的概念、产生原因及处理目的出发,分别给出正态分布、极值分布、威布尔分布下异常值的诊断方法与应用;针对检测结果,概括了对异常值进行平滑处理的基本技术。



第六章为“统计数据的变换处理”。以统计数据变换处理的作用为导向,阐述包括指标值域变换、逆变换、属性构造等在内的各种数据变换方法的基本原理,并给出数据变换方法的典型示例。

第七章为“统计数据的平衡协调性检验”。从相关指标的验证核查、经验参数比较、变动趋势对比、因果关系检验、支撑度判断和反常结果判断等几个方面全面阐释了统计数据平衡协调检验的具体方法及应用。

第八章为“国际上统计质量管理总体架构”。国际上一些国际组织及政府统计机构以“官方统计基本原则”为基础,从数据生产视角提出统计活动的基本准则,从数据发布视角提出统计资料的基本要求,并针对数据质量问题给出许多单项质量评估框架和综合性质量保证体系。

第九章为“三种主要国际统计质量管理框架比较”。对国际货币基金组织、欧盟统计局及经合组织的统计质量管理框架,在质量维度与含义、整体框架、评估方法与工具等方面的区别与联系进行详细对比分析,从中得到对我国统计数据质量管理的若干启示。

第十章为“国家统计质量保证框架、模板及导则”。针对联合国统计委员会第41届会议上的国家统计质量保证框架议题,介绍了国家统计质量保证框架的产生背景、主要内容,概括国家统计质量保证框架模板的主要结构及导则,阐述国家统计质量保证框架的作用。

第十一章为“基于利益相关者视角的统计数据质量评估与管理”。由于统计数据质量涉及数据的生产者、使用者和提供者等直接参与主体及众多的外部监督主体,这一章基于利益相关者理论开展统计数据质量利益相关者的识别、分类,给出利益相关者视角的数据质量评估方法及数据质量管理策略。

第十二章为“结语”。总结前文分析得到的主要结论,概括本书的创新与不足,展望未来研究趋向,并从数据预处理、利益相关者、国际统计质量框架、法制保障、思想观念、管理体制、方法制度、基础建设、全面管理等方面提出提升统计数据质量的对策措施。

上篇基于数据预处理视角,把涉及统计数据质量诊断的诸多方



法纳入统一分析框架,并利用这些方法开展应用分析;下篇以国际上统计数据质量管理框架为基础,基于利益相关者视角提出统计数据质量的评估方法及管理策略。本书的研究对于开展统计数据质量诊断与管理具有重要的理论与实践价值。

理论方面:(1)研究视角的创新。抛开一般从数据来源及数据收集过程来考虑提高统计数据质量的常规思路,抓住数据收集之后、数据分析之前这一间隙,提出利用数据预处理方法开展数据质量的诊断这一独特视角。另外,不再单纯从统计数据生产者即政府统计机构的角度来考虑提高数据质量,而是依据利益相关者理论,从生产者、使用者、提供者及监督者的整合视角,全方位开展统计数据质量评估与管理。(2)系统地提出数据预处理的方法体系和研究框架。将六大类数据预处理方法纳入到一个系统框架当中,并与数据预处理的四大过程对应起来,阐述每一类方法的基本原理和实际应用。(3)拓展了数据预处理的内涵。数据预处理原本是数据挖掘中的一个步骤,本书对其内容进行较大的扩展,使其具有更为丰富的内涵。(4)质量管理理念的拓展。将操作层面的数据质量评估方法与战略层面的数据质量管理相结合,由单一面向统计产品的质量评估扩展为将统计过程与统计产品相结合的全面质量管理。

实践方面:(1)提出一种诊断数据质量的有效策略。面对日益突出的数据质量问题,另辟蹊径地进行数据预处理,不失为提高数据质量,保证数据分析结果客观性、真实性的有效途径。(2)为社会各界开展数据质量诊断及评估提供了一套有力工具。结合六大类数据预处理方法,根据不同的数据类型和不同要求,可以有针对性地选择恰当方法对数据进行预处理,评估数据质量。(3)拓宽了提高统计数据质量和统计公信力的途径。面对日益突出的数据质量问题,让所有的利益相关者都参与到统计数据质量评估与管理中来,对于提高社会各界对统计数据的认同度,提升统计公信力具有显著的现实意义。(4)发掘了国际上统计质量保证框架对我国统计数据质量管理的借鉴价值。本书对国际组织和不同国家的统计质量保证框架进行系统的综合、比较,从中提炼出开展统计数据质量管理的有益经验,对于



我国各级政府统计机构开展实际的统计数据质量管理,实施有效的对策与措施,具有启示意义。

由于统计数据预处理涉及的方法较为庞杂,本书对其中一些方法未能展开详细论述,譬如缺失值处理的插补方法、时序数据及面板数据的异常值诊断等;另外,对于在统计数据质量诊断中如何恰当、有效地选用数据预处理方法,也只是做出了初步尝试,尚需全面、系统地加以梳理和整合。基于利益相关者视角的统计数据质量评估与管理,只是提出了初步的评估方法、组织形式及管理策略,未能开展实际的统计数据质量评估,各种评估方法的有效性还有待实证检验。这些不足有待今后在追踪国内外最新研究成果的基础上,进一步开展深入探讨。

目 录

第一章 总 论	1
一、中国统计数据质量的论争	1
二、统计数据质量研究综述	5
三、统计数据质量的内涵.....	17
四、统计数据质量问题产生的原因.....	24

上篇 统计数据质量诊断

第二章 统计数据质量诊断的新视角——数据预处理	31
一、统计数据质量诊断的必要性.....	31
二、统计数据质量诊断的新视角.....	33
三、统计数据预处理的过程.....	34
四、统计数据预处理的方法体系.....	36
第三章 统计数据的初步审查	42
一、描述及探索性分析.....	42
二、信度和效度检测.....	47
第四章 统计数据的缺失值处理	62
一、缺失数据的机制和模式.....	63
二、缺失数据的事前处理方法.....	65
三、缺失数据的事后处理方法.....	69



第五章 统计数据的异常值处理	83
一、异常值诊断概述	83
二、正态分布下异常值诊断及应用	88
三、极值分布下异常值诊断及应用	98
四、威布尔分布下异常值诊断及应用	100
五、异常值的平滑处理	111
第六章 统计数据的变换处理	114
一、统计数据变换的作用	114
二、统计数据的变换方法	115
三、数据变换方法的应用	121
第七章 统计数据的平衡协调性检验	126
一、相关指标的核查验证	126
二、相关指标的经验参数比较	128
三、相关指标的变动趋势对比	130
四、相关指标的因果关系检验	132
五、相关指标的支撑度判断	136
六、统计数据的反常结果判断	137
七、其他手段	139
下篇 统计数据质量管理	
第八章 国际上统计数据质量管理总体架构	143
一、国际上统计数据质量管理体系	143
二、统计数据质量管理的最基本原则	145
三、统计数据“生产”的基本准则	146
四、统计数据“发布”的基本要求	149
五、统计数据质量的评估与管理框架	150
六、本章小结	154



第九章 三种主要国际统计数据质量管理框架比较	155
一、三种主要国际统计质量管理框架概述	156
二、质量维度及含义比较	158
三、整体框架的比较	161
四、质量评估方法与工具的比较	165
五、本章小结	167
第十章 国家统计质量保证框架、模板及导则	169
一、国家统计质量保证框架的产生背景	169
二、国家统计质量保证框架的主要内容	172
三、国家统计质量保证框架模板及导则	174
四、国家统计质量保证框架的作用	176
第十一章 基于利益相关者视角的统计数据质量管理	177
一、统计数据质量的利益相关者识别	178
二、利益相关者视角的统计数据质量维度	185
三、利益相关者视角的统计数据质量评估	187
四、利益相关者视角的统计数据质量管理策略	196
五、本章小结	199
第十二章 结 语	201
一、主要结论	201
二、未来展望	205
三、对策措施	208
参考文献	216
附 录	229
后 记	241



表目录

表 1-1	不同国家或国际组织的统计数据质量定义	/19
表 3-1	用交叉表寻找逻辑上的不一致	/46
表 3-2	对 10 名学生的两次调查结果	/49
表 3-3	对 10 名学生两次测量结果的相关系数	/49
表 3-4	同一组消费者在不同时间对 10 种商品质量的排序	/50
表 3-5	对某一对象两次观察结果	/50
表 3-6	对 203 人的两次调查结果	/51
表 3-7	复本调查结果	/52
表 3-8	皮尔逊相关系数	/52
表 3-9	对 5 位学生随堂测验结果	/53
表 3-10	对 6 名学生的随堂测验结果	/55
表 3-11	项目总计统计量	/55
表 3-12	α 信度系数	/55
表 3-13	各分项得分与总分之间的相关系数	/57
表 3-14	因子提取结果	/59
表 3-15	旋转后的载荷矩阵	/61
表 4-1	变量 X_{miss} 和 Y_{miss} 的相关系数	/77
表 4-2	EM Correlations ^a	/78
表 4-3	EM Correlations ^a	/79
表 4-4	几种插补方法的结果比较	/82
表 5-1	$F'=1$ 时的 G'	/93
表 5-2	$D'=1, G'>1$ 时的 Z 值	/94
表 5-3	1990—2007 年中国城镇居民人均可支配收入和人均消费支出数据	/94
表 5-4	p_i (假定 $C_{1995}=3600$ 元)	/97
表 5-5	p_i (假设 $C_{1995}=3630$ 元)	/97
表 5-6	T' 分布的分位数 ($\hat{\sigma}$ 是 σ 的 BLUE 估计)	/108

表 5-7	T' 分布的分位数($\hat{\sigma}$ 是 σ 的 BLIE 估计)	/108
表 6-1	同学的三门课程成绩	/122
表 6-2	同学三门课程成绩的标准分	/122
表 6-3	2006 年各省城市化与经济增长的协调度	/124
表 7-1	序列平稳性检验结果	/134
表 7-2	Johanson 协整检验结果	/134
表 7-3	财政总收入与经济增长的因果关系检验结果	/135
表 7-4	地方财政收入与经济增长的因果关系检验结果	/136
表 8-1	官方统计基本原则	/145
表 9-1	三种统计质量保证框架的质量维度比较	/160
表 9-2	IMF 的数据质量评估框架结构	/161
表 9-3	OECD 的统计质量保证框架	/164
表 11-1	统计数据质量利益相关者影响/利益矩阵	/185
表 11-2	利益相关者视角的统计数据质量维度、评估方法及指标	/189



图目录

- 图 2-1 基于数据预处理视角的统计数据质量诊断框架 /34
- 图 2-2 数据预处理的过程 /35
- 图 2-3 数据预处理过程与方法的关联 /37
- 图 3-1 大学生心理测量数据结构 /57
- 图 3-2 战略绩效控制与绩效测评指标说明 /59
- 图 4-1 变量 X 与 Y 的散点图 /77
- 图 4-2 变量 Y 缺失值的均值替换效果 /78
- 图 4-3 变量 Y 缺失值的回归替换效果 /79
- 图 8-1 国际上统计数据质量管理体系的总体架构 /144
- 图 9-1 ESS 的统计质量管理框架及要素间的关系 /163
- 图 11-1 统计数据质量的利益相关者 /182

第一章

总 论

无论是宏观经济决策还是微观企业经营,都需通过观察和分析数据而作出相应的决策。社会经济发展需要的是真实、准确、及时、全面的统计数据,没有质量保证的数据只是一堆数字垃圾,不仅毫无用处,而且可能形成误导。随着经济、社会发展和统计知识的日益普及,党和政府对统计数据质量的要求越来越高,社会各界对统计数据问题也越来越加以关注。统计数据质量的优劣不仅直接影响统计部门的形象,而且影响国民经济管理和政府决策,关系到企业生产经营活动与人民生活,甚至一定程度上反映出一个国家的管理水平^①。

一、中国统计数据质量的论争

中国统计数据的质量状况一直受到国内外的广泛关注, Tom Orlik(2009)的“如何看待中国经济减速时期的 GDP 统计”一文再次引起人们对中国 GDP 数据的关注^②;路透社网站出现 Wheatley (2009)一篇题为“中国统计数据迷雾重重 经济学家不知所措”的评论,引述了多位西方和我国学者对中国统计数据的观点,大多认为中

① 陈建成、庞新生:《统计分析理论与方法》,中国林业出版社 2006 年版。

② Tom Orlik:《如何看待中国经济减速时期的 GDP 统计》,《华尔街日报》,2009 年 4 月 10 日。