

政府统计数据 质量体系研究

ZHI LIANG TI XI YAN JIU

傅德印 著



甘肃人民出版社

政府统计数据 质量体系研究

傅德印 著

甘肃人民出版社

责任编辑:刘新田

封面设计:贾 雯

政府统计数据质量体系研究

傅德印 著

甘肃人民出版社出版发行

(730000 兰州市滨河东路 290 号)

各地新华书店经销 武威市印刷厂印刷

开本 850×1168 毫米 1/32 印张 7.5 插页 2 字数 179 千

2000 年 1 月第 1 版 2000 年 1 月第 1 次印刷

印数:1—1,300

ISBN 7-226-02157-9/C·187 定价 18.00 元

ABSTRACT

The official statistical data quality management system(OSDQMS)is very important for both data users and data producers. The purpose of this paper is to combine the technique of official data quality control with the organization management of statistical process, establish the frame of OSDQMS. And probe into the implementing and carrying out of this frame, in the light of the needs of statistical information sector and the international standard of quality management. On the basis of defining the official statistical data quality, analyzing the present situation of official statistical data quality and quality management in China.

First this paper reviews the present situation of study of official statistical data quality management and control. Then in the part one, the paper defines the official statistical data quality, analyses the present situation of official statistical data quality and quality management in China. And considers the OSDQMS includes two main aspects, they are official statistical data quality control system(OSDQCS) and official statistical data quality assurance system(OSDQAS).

In the part two, this paper probes into the principle and structure of the OSDQCS from three main aspects. They are list approaches, Comparative analysis approaches and error model approaches And probes into the application of FDA and discriminate analysis to the control of statistical data quality.

In the part three, this paper probes into the principle and structure of the OSDQAS. It considers that the key of OSDQAS is the leadership responsibility, personnel and resources, and structure of the quality assurance system that include quality loop, work elements, quality documentation and records, internal quality audit etc. In this part, some recommendations for impenenting and establishing the OSDQAS will be put forward, they concern with integrate design, the edition of documentation and the operation of the OSDQAS.

Finally in the part four, the paper probes into the application of OSDQMS to population census and CPI data quality management. And probes into the application of OSDQMS to the control of false statistical data.

序

袁 卫

政府统计数据质量关系着政府决策的科学性和正确性,关系着统计工作的声誉和生命,特别是在信息技术数字化、网络化和经济发展全球化的未来社会中,构成信息主要内容的统计数据质量更为重要。长期以来,如何提高政府统计数据质量一直是困扰着国内外政府统计机构的重大难题,也一直是我国政府统计工作的首要任务。提高和保证统计数据质量,仅仅强调数据的准确性是远远不能满足数据用户多方面需求的,要彻底地解决好数据质量这一问题也不是靠几次执法大检查等临时性措施所能办到的,这就有必要根据政府统计信息产业发展的要求,在正确的数据质量观念基础上,从理论和实践上研究探讨数据质量管理体系的建立和实施问题。

目前,国外关于统计数据质量的研究经历了三个阶段。第一阶段是从本世纪初30年代至40年代,主要是围绕着统计数据准确性,样本代表性问题,沿着统计数据准确性与抽样误差,统计数据准确性与非抽样误差两条思路展开的。第二阶段是50年代初至70年代,主要是围绕着统计调查误差模型的建立,以及敏感性问题的随机化方法等专项问题研究展开的。第三阶段始于70年代末至现在,这一时期统计数据质量特性由准确性发展到全面性、系统性、一致性、目的性、及时性、保密性、经济性等比较宽泛

的内容,对统计数据质量问题的研究也发展到如何建立有效保证、控制、评估体系上来。

相对而言,国内关于统计数据质量问题的研究还比较薄弱。对统计数据质量涵义还缺乏相当的共识,关于非抽样误差的研究才是近几年的事,对统计数据质量“全面质量管理”还局限于几项重大的普查应用中,强调数据的准确性多,对其他质量特性强调的少,对数据质量的事后检查评估多,而对事前的预防控制少。因此,结合我国政府统计的实际情况,借鉴国外的一些研究经验,对政府统计数据质量体系进行探讨,是一件十分有意义的事情。

傅德印同志近几年一直从事政府统计数据质量问题的研究,本书是在他的博士学位论文的基础上修改而成的。这本专著具有以下几个特点:

1. 比较全面、客观、多视角地阐述了政府统计数据质量概念,较全面地总结了我国政府统计数据质量研究及管理经验。

2. 首次把 ISO9000 系列标准的概念引入到政府数据质量控制中,提出统计数据质量管理体系是完善统计运行体系的重要组成部分的观点,并把政府统计数据质量控制技术与工作过程的组织管理进行结合,探讨政府统计数据质量管理体系的框架、建立、实施和运行。提出统计数据质量管理体系的主体框架包括政府统计数据质量控制技术体系和政府统计数据质量保证体系。

3. 比较深入地分析了统计数据质量控制技术体系的内容结构,即分类控制与评估技术,比较分析及探索性数据分析技术,误差模型的建立与分析技术,并对探索性数据分析,贝叶斯逐步判别分析等方法在数据质量控制中的应用进行了尝试和研究。

4. 比较深入地分析了统计数据质量保证体系的内容结构、建立、实施和运行。结合国际标准,提出了统计数据质量体系中质量环的 7 个环节、15 个作业要素、以及控制程序。

5. 理论研究与应用研究相结合, 是该文的方法论特点。运用统计数据质量管理体系, 对我国人口普查和居民消费价格指数数据质量管理体系的建立与实施进行了应用研究, 并就统计数字中的虚假问题提出了解决措施。文中关于无回答误差模型、测量误差模型的论述, 用概率预报的方式公布统计资料, 编制政府统计数据质量保证体系文件等观点和建议, 也都值得一读。

傅德印同志能够结合政府统计实际, 就政府统计数据质量管理体系进行比较全面、系统的探索, 是很可贵的。当然, 这其中的许多观点和方法还有待于进一步完善, 其在实际工作中的可操作性, 也有待于今后实践的检验。我之所以愿意把本书推荐给关注政府统计数据质量控制、评估方法问题研究的有关专家和读者, 原因之一在于希望能够引起更多的理论研究人员和实际工作者的兴趣来致力于该领域的进一步研究, 把我国政府统计数据质量提高到新的水平。

1999 年 10 月于中国人民大学

目 录

第一章 导 论	(1)
1.1 政府统计数据质量研究综述	(1)
1.2 本书出发点、目的及结构安排	(12)

第一部分 政府统计数据质量管理体系原理与框架

第二章 政府统计数据质量管理体系的构建	(15)
2.1 政府统计数据质量涵义界定	(15)
2.2 我国政府统计数据质量及其管理的总体认识	(22)
2.3 政府统计数据质量管理体系原理与框架	(28)

第二部分 政府统计数据质量控制技术体系

第三章 分类控制与评估技术	(39)
3.1 分类控制与评估技术一般原理	(39)
3.2 分类控制与评估技术体系的构建及应用	(40)
3.3 对“误差文档及监测方法”的借鉴应用	(49)
第四章 比较分析与探索性数据分析技术	(57)
4.1 比较分析的一般原理	(57)
4.2 比较分析技术体系的构建及应用	(58)
4.3 探索性数据分析技术的应用	(69)
第五章 误差模型建立与分析技术	(81)
5.1 误差模型建立与分析技术的一般原理	(81)
5.2 无回答误差模型技术的应用	(82)

5.3 统计测量误差模型技术的应用	(97)
5.4 判别分析方法在汇总数据质量控制的应用	(107)

第三部分 贯彻国际标准,建立健全政府 统计数据质量保证体系

第六章 政府统计数据保证体系的原理、结构和 主要要素	(117)
6.1 政府统计数据质量保证体系的研究现状	(117)
6.2 政府统计数据质量保证体系的原理	(122)
6.3 政府统计数据质量保证体系的结构	(126)
6.4 政府统计数据质量保证体系的作业要素 及控制	(130)
第七章 政府统计数据质量保证体系的建立、实 施及运行	(141)
7.1 政府统计数据质量保证体系的总体设计	(141)
7.2 政府统计数据质量保证体系文件的编制	(145)
7.3 政府统计数据质量保证体系的运行	(150)
7.4 数据质量保证体系建立实施中可能的问 题及出路	(157)

第四部分 政府统计数据质量管理体系应用探讨

第八章 建立健全我国人口普查统计数据质量管 理体系的思考	(161)
8.1 建立健全人口普查统计数据质量控制技 术体系	(162)
8.2 建立健全人口统计数据质量保证体系	(177)
第九章 建立健全我国居民消费价格指数质量管 理体系的思考	(186)

9.1	如何看待我国目前居民消费价格指数的数据质量	(186)
9.2	建立健全我国居民消费价格指数的质量控制技术体系	(193)
9.3	建立实施我国居民消费价格指数质量保证体系	(195)
第十章	政府统计数据质量管理体系与统计中虚假问题的控制	(200)
10.1	政府统计中虚假数据的表现	(200)
10.2	虚假数据的危害及影响	(204)
10.3	虚假数据的原因分析	(206)
10.4	虚假数据的控制与政府统计数据质量管理体系完善	(209)
附表 1.	原始数据茎叶图表(1993)	(216)
附表 2.	原始数据茎叶图表(1994)	(218)
附表 3.	回判结果表	(220)
附表 4.	对待判别样品的分类结果	(221)
后 记		(222)

第一章 · 导 论

1.1 对政府统计数据质量研究的综述

1.1.1 国外关于政府统计数据质量研究的进程

对政府统计数据质量的研究是随着社会、经济、科技的发展对大规模的统计数据搜集系统的需求而产生并不断发展的。而历史上,人们对统计数据误差理论的研究早已展开。最早可首先追溯到1812年拉普拉斯(P.S.de Laplace)出版的《概率分析理论》一书中,就建立了观测误差的理论。其次是高斯(C.F.Gauss)提出了误差分布曲线即正态分布曲线,奠定了误差理论的基础。另外再如泊松(S.D.B.Poisson)应用概率研究人口统计、司法统计、医疗卫生统计中的统计误差方面的问题^①。但是限于当时的历史条件,政府大规模的统计数据搜集系统还不够发达,因此这些统计学家对统计误差的研究主要集中在理论上以及在生产、科学实验中对测量数据误差的研究。

随着政府统计的发展,特别是随着经济、社会、科技发展对政府大规模统计数据搜集系统的需求,从本世纪初,人们不仅研究误差理论,测量误差,而且开始研究统计调查误差。不仅研究调查数据误差,而且研究统计数据质量,以及政府统计数据质量。具体可

① 周概容:《统计学辞典》,中国统计出版社1993年6月第1版第415页。

划分为以下几个阶段:

第一阶段(本世纪初~三、四十年代),该时期统计数据质量研究主要是围绕着调查数据的准确性、样本的代表性,沿着两条思路展开的。

一是对调查数据准确性与抽样误差的研究。本世纪初期正值抽样调查与普查展开激烈争论的时期。为此围绕统计数据质量,人们首先讨论的是,调查结果的准确性是取决于观察数量的多少,还是取决于正确“代表性”方法(抽样方法)的问题。以挪威统计学家凯尔(A. N. Kiar)为代表提出“代表性”方法,而以迈耶(J. V. Mayer)教授等为代表则反对用抽样方法,坚持普查方法不容动摇的观点。直到1903年柏林国际统计学会大会之后,抽样方法才得到大多数统计学家的认可,并且日益得到应用和普及。^①

抽样调查方法的地位确立后,围绕统计数据质量人们讨论的是,调查数据的准确性关键又取决于样本的代表性如何,于是又产生了有目的抽选代表性样本(如目的抽样、定额抽样等)与随机抽选样本两种意见。以英国统计学家鲍莱(A. L. Bowley)为代表主张采用随机抽样方法,补充了凯尔的创始工作。随机抽样的科学地位确立后,人们研究的是各种随机抽样方法的设计及不同设计方法的抽样误差的计算和控制。期间最有影响以及给随机抽样带来最大变化的是奈曼(J. Neyman)1934年关于分层抽样最优分配及回归、比率估计的研究,对各种抽样设计的抽样误差进行了有效估计,产生了今天所谓的“可测的设计”^②。统计学家们把奈曼的研

① You Poh Seng(1951). Historical Survey of the Development of Sampling Theories and Practice. Journal of Royal Statistical Society, series A, 114, 214~231.

② [美]胡善庆著,张尧庭等译,《抽样调查一百年》,《抽样调查理论与实践》,中国统计出版社1995年第1版第1页。

究视为抽样方法历史的一个转折点。实际上,这也是统计数据质量研究的一个重要里程碑。

另一条思路是围绕着统计数据的准确性与非抽样误差。^①本世纪初人们就开始注意到非抽样误差的影响。1915年鲍莱在进行一项有关就业和贫困问题的调查时就指出,影响调查结果的非抽样误差因素包括,获得的资料不正确、不真实和贫困定义不严格、不确切两方面。第一次世界大战后,实际工作者们在许多应用中都发现各种非抽样误差如测量误差、无回答误差等对全部调查误差的影响比抽样误差的影响更重要、更明显,这就激发了人们对非抽样误差的研究和讨论。而这种研究实际上是随着政府进行大规模统计数据搜集而展开的,并在30年代~40年代取得了重大的进展。当时最有影响的是美国国情普查局进行人口普查时对非抽样误差的研究,以及印度政府进行粮食产量估计时对非抽样误差的研究。其中美国国情普查局对非抽样误差的研究主要包括:(1)在普查报告中提出统计资料产生矛盾和误差的可能原因;(2)对普查数据的误差及准确性进行分析;(3)研究如何利用一种“单独复核的方法”来计算误差。^②复核的内容包括,调查范围误差,如人口、住户和农场的遗漏,重复错列的误差,填报内容误差,如调查表中具体问题的漏报或误报等。

第二阶段(50年代初~70年代),该时期政府统计数据质量的研究是围绕着统计调查误差模型的建立及由此展开的专项研究展开的。调查误差模型是以整体的方式测定抽样误差和非抽样误差

① 金勇进编著,《非抽样误差分析》,中国统计出版社1996年2月第1版第3页。

② E.S.麦克司,W.P.马丁,H.尼塞孙:《论美国1950年人口普查的复查——关于调查设计的一个专题研究》,统计译文专辑(11),统计出版社1958年版。原载《美国统计学会会刊》1953年6月号。

的影响。对调查模型的研究最早始于 40 年代末,1946 年汉森(M. H. Hansen)和赫维茨(W. N. Hurwitz)提出了无回答问题。^①50 年代研究取得成果。60 年代由汉森和赫维茨等首次提出完整的调查误差模型。^②即

均方误差 = 测量误差 + 抽样误差 + 测量误差和抽样离差的协方差 + 偏差的平方

该模型的建立,对调查过程中误差的控制、评估奠定了重要的分析基础,一方面,该模型利用均方误差的概念把总误差分解为各个相对重要的误差组成部分,并使测定每部分在总误差中的影响成为可能;另一方面,在此基础上,人们还可以进一步研究每一分量的分解及测算。也正因为如此,六七十年代后,围绕调查误差的许多研究是要找出均方误差中占较大份额的组成部分,并研究减少这些成分的方法。如对子样本重复计量、交叉子样本的方法以及二者相结合的方法。^③同时这一时期也出现一些专项研究,最典型的是 60 年代沃纳(S. L. Warner)提出对敏感性问题调查的随机化方法。^④

第三阶段(70 年代末至现在),该时期统计数据质量的含义扩大,对统计数据质量的研究发展到如何建立有效的保证、控制、评估体系。

① M. H. Hansen and W. N. Hurwitz (1946). The problem of Non-response in Sample Surveys, Journal of the American Statistical Association, 41, 516 ~ 529.

② M. H. Hansen and W. N. Hurwitz, and Bershad, M. (1961). Measurement errors in censuses and surveys. Bulletin of the ISI, 38. 2. 359 ~ 374.

③ 科克伦(1977)著,张尧庭等译,《抽样技术》,中国统计出版社 1985 年 4 月第 1 版第 588 页。

④ S. L. Warner(1965)Randomized response:A Survey Techniqu for Eliminating Evasive answer bias. Journal of the American Statistical Association 60, 63 ~ 69.

政府大规模统计数据搜集系统的日益制度化,抽样调查地位的确立及调查模型的建立,使对统计数据质量的研究趋势是研究调查设计的规范化、调查误差与调查费用的平衡,而且随着调查内容的复杂化,对被调查对象提供资料的保密及个人隐私的保密也已成为统计调查设计质量考虑的重要内容。因此统计数据质量的含义得到扩展,由单纯的统计数据准确性扩展到比数据准确性更为丰富的内涵。戴伦纽斯(T. Dalenius)提出统计数据质量的“测量向量”^①,包括统计数据的准确性、经济性、保密性、相关性、时效性、数据详细程度等。1980年联合国统计局出版的《统计组织手册》中,对官方统计资料提出八项要求,^②实际上也远远超出统计数据准确性、及时性的要求,进一步明确了政府统计数据质量的综合含义。

70年代石油危机后,使国外政府非常关心统计数据质量对决策的影响,信息产业的发展,更促使人们开始研究利用现有的各种知识,综合利用,对政府统计数据质量进行综合的控制。戴伦纽斯提出数据质量的综合控制应包括调查设计控制和调查实施控制两部分。调查实施控制又分为预防控制和过程控制。80年代初美国的许多政府统计机构以及大的公司对本单位的统计数据质量都进行了综合质量控制,1988年4月美国田纳西 Oak Ridge 统计数据质量专题研讨会上,提出了统计数据质量环的一般内容并进行讨论。而且就统计数据质量控制达成以下共识^③:(1)简单的质量控制应在实施更困难、花费更高的步骤发生之前进行,即强调统计数

① T. Dalenius(1983). Errors and other limitations of survey, Statistical Methods and the Improvement of Data quality, Edited by T. Wright, Academic Press Inc.

② United Nations, Handbook of Statistical Organization, New York, 1980.

③ Data Quality Control—Theory and Pragmatics, Edited by Gunar F. Liepins and V. R. R. Uppuluri, Marcel Dekker, INC. 1990.

据质量的事前控制。(2)数据误差和偏差应在质量控制环最早的可行阶段中发现,即强调统计设计质量。(3)软件应标准化,即强调统计数据质量管理方法体系的标准化。(4)进一步发展统计数据质量的综合测量和对用户提供统计数据使用指南。这实际上是为进一步的规范化的数据质量控制提出了一个管理模式或标准。此后,不仅美国,而且世界上许多其他国家如意大利^①,瑞典^②的农业普查中都进行了统计数据质量的综合控制。相应地,如何进行统计数据质量综合控制也成为重要的研究课题。

1.1.2 目前国外关于统计数据质量研究的主要内容

纵观国外目前对统计数据质量的研究,不难发现,其研究内容是极其广泛的,具体集中于^{③④}:

(1)误差的来源、分类以及如何对误差文档及监测的研究。误差分布以及误差模型的研究。

(2)统计调查设计质量与规范的研究(如问卷设计、调查规模、分组体系、保密问题)。对统计数据搜集目的的研究。用户的最初要求,潜在用户及潜在需求的研究。调查实施计划框架充分性的研究。对调查过程的综合概括以及对统计数据质量环的研究。

(3)统计数据搜集质量的研究(如回答者误差、调查员误差、调查方式的影响、数据搜集质量的测量、新技术对统计调查及其数据

① Gualtiero M. F. Schirinzi(1994), Quality Control in Italian agricultural censuses, Statistical Journal of United Nations ECE11 1~17.

② Knut Medin, Quality Assurance and quality assessment in official statistics: the case of Sweden's panel census of Agriculture, Statistical Journal of United Nations ECE11 1~17.

③ Statistical Methods and the Improvement of Data quality, Edited by T. Wright, Academic Press Inc.

④ 《Statistical Journal of United Nations ECE.11》,1994.

质量的影响)。数据搜集系统的问题及处理。普查抽样调查有机结合的研究。

(4)统计数据处理及操作质量的研究(如编码、误差手册、自动编码、数据录入质量、数据编辑、数据处理质量及改进、数据处理过程的控制系统)。

(5)质量评估和控制的研究(如调查质量的测量、数据质量评估方法、调查误差模型及分析、误差文档及监测、用户的质量评价、质量管理)。数据质量的测量与控制方法的研究。统计数据报告内容及结构的研究。

(6)数据误差对参数估计、统计分析的影响。调查误差对数据质量的影响及分析。参数估计的准确性及精确性。探索性数据分析和模型识别技术的应用。奇异值诊断方法和数据编辑技术的应用。不完整数据处理计划和实施。

(7)统计数据质量与调查费用及调查效率的研究。

(8)统计数据质量改进中的问题及对策研究。提高统计数据质量的管理步骤。统计理论与实际工作者如何提供支持管理者决策的统计数据。统计数据内部一致性,外部一致性的检验。统计数据文件、统计资料的管理。

(9)目前统计数据质量控制的成功经验。目前统计数据的质量水平。统计数据质量理论研究成果与统计数据质量改进实践相结合的研究。

显然,从一般的误差理论,到调查数据的误差理论;从抽样误差到非抽样误差;从单纯的数据误差与准确性到数据的综合质量;从民间统计数据质量到政府统计数据质量;从统计数据质量的事后检验、控制到统计数据质量的事前保证;从统计数据质量的保证与控制技术到政府统计数据质量的管理,从理论到实践等都有研究。