北京联合大学
服务外包人才培养模式创新实验区
专项资助

问卷调查数据分析实务

[张士玉 著]

首都经济贸易大学出版社
Capital University of Economics and Business Press

北京联合大学
服务外包人才培养模式创新实验区
专项资助

问卷调查数据分析实务

[张士玉 著]

首都经济贸易大学出版社
Capital University of Economics and Business Press
·北 京·

图书在版编目(CIP)数据

问卷调查数据分析实务/张士玉著. —北京:首都经济贸易大学出版社,2015.2

ISBN 978-7-5638-2291-1

Ⅰ.①问… Ⅱ.①张… Ⅲ.①问卷调查—统计分析 Ⅳ.①C915-03

中国版本图书馆 CIP 数据核字(2014)第 239145 号

问卷调查数据分析实务

张士玉 著

出版发行	首都经济贸易大学出版社
地　　址	北京市朝阳区红庙(邮编 100026)
电　　话	(010)65976483　65065761　65071505(传真)
网　　址	http://www.sjmcb.com
E-mail	publish@cueb.edu.cn
经　　销	全国新华书店
照　　排	首都经济贸易大学出版社激光照排服务部
印　　刷	北京京华虎彩印刷有限公司
开　　本	710 毫米×1000 毫米　1/16
字　　数	206 千字
印　　张	11.5
版　　次	2015 年 2 月第 1 版　2015 年 2 月第 1 次印刷
书　　号	ISBN 978-7-5638-2291-1/C·116
定　　价	26.00 元

前　言

问卷调查和数据分析是当今服务外包领域的重要组成部分，同时也是社会科学研究的重要方法。数据分析工作质量的高低，不仅关系到研究者的研究目的是否可以达到、研究成果是否令人满意，而且关系到社会资源是否被有效利用的问题。一份调查问卷发出后，少则需要几百人，多则数千数万人，全国人口1%抽样调查则需要上千万人填写数据。一次社会调查的完成，不仅仅涉及许多人的劳动，更重要的是倾注了被调查者的诚实和精力。所以，一套调查数据是珍贵的社会资源，如果不对其进行高质量的分析和深入的挖掘，是对社会资源的浪费，是对被调查者的诚实和精力的不尊重，更是对自己所研究项目的不负责任。常常听调查者抱怨说：问卷难发、被调查者不认真等。但是反过来调查者有没有扪心自问：我们是否珍惜了宝贵的数据资源？是否可以高质量地分析数据？是否可以发现有价值的数据规律？笔者常见许多专业公司或研究者辛辛苦苦完成了调查后，对于数据分析却比较简单，看到高质量的数据没有被充分地分析和挖掘，感到十分可惜。

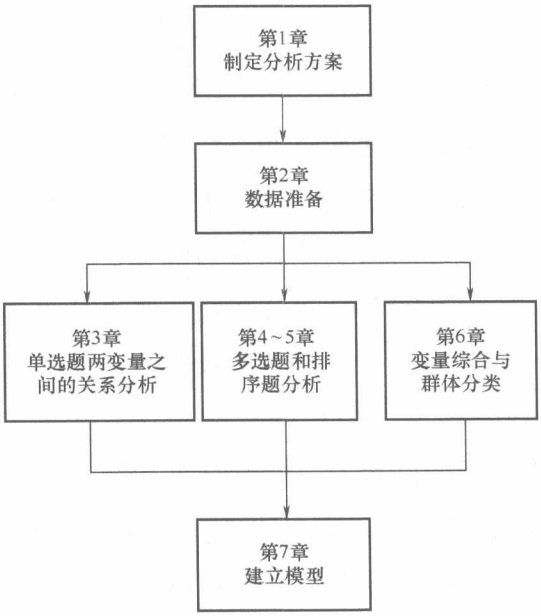
在此首先要明确的问题是，本书是基于撰写数据分析报告而非专题论文，两者具有紧密关系但却有很大不同。第一，数据分析报告要展示数据特点并发现隐藏在数据背后的规律，其要求具有全面性，不要遗漏重要特点和规律，所以要多视角看问题、采用多种方法、对多变量进行组合尝试分析，许多或大量的工作可能都是白做，只有部分结果具有价值；论文只是对其中某一个具有价值的结果进行深入论述，主题明确。第二，数据分析报告以定量分析为主，定性解释为

辅;论文则是以论据支持论点,定性论述的比重较大。一篇好的论文可能是基于好的分析报告,一篇好的分析报告则可以形成若干篇视角不同的论文。从这个意义上来说,报告是论文的前期工作和基础。

目前所见数据分析的书籍众多,主要分为三大类:一是从方法论出发,分为统计学、数据挖掘等;二是从工具使用出发,最为典型的是社会统计软件(SPSS);三是从某类模型出发。

本书旨在从研究目的出发,结合变量类型选择方法和工具,以系统的观点,从整体上考察各个变量及其关系,再考虑采用方法和工具,从而避免数据分析中常见的问题,帮助读者写出高质量的、分析深入的数据分析报告。

本书结构如下:



对调查数据的综合分析实际上是一个博大精深的领域,一份好的数据分析报告是科学与艺术的结晶。本书只是希望以此皮毛内容起到抛砖引玉之作用。

在此，笔者首先要感谢北京联合大学对学术研究的重视与导向，以及对本书给予的直接资助。同时，在本书的形成过程中笔者直接或间接地受到过许许多多人的帮助：对外经济贸易大学郝旭光教授所主持的国家社会科学基金项目提供了数据支持和实验领域；北京联合大学顾志良教授、何勤副教授、陶秋燕教授、陈琳教授、王晓红教授等所主持的项目都提供过数据支持和实验应用领域；马丽仪副教授在熵学方法方面给予了建议；招生就业处高桥处长、张伟处长所提供的学生调查数据提供了数据支持；刘文芝博士校对了全书。还要感谢许许多多的同行学者，包括参考文献中提到和未提到的学者对本人的启示和在数据分析领域中做出的贡献。

本书的形成少不了家人的支持和帮助，首先感谢父母的养育之恩；感谢妻子张肖女士的理解与支持，是她在厨房的辛苦换来了我在书房的专注；感谢儿子张然，是他的自立能力使我免去后顾之忧。

01
CONTENTS

目 录

02
CONTENTS

1 制定分析方案

本章首先在引言中论述综合采用不同研究方法的重要性和必要性，然后归纳了在调查数据分析的实践中所面临的主要问题，指出产生问题的原因，进而提出以系统的观念、从所研究的目标出发、多视角多方法地分析数据，提出了在实践中要重视分析方案的设计工作，并通过实例加以说明。

1.1 引言

在以科学方法追求真理的过程中，逻辑与证据两者相辅相成，缺一不可，陈晓萍等(2008)在《组织与管理研究的实证方法》一书中明确指出："没有数据的逻辑或没有逻辑的数据只是科学方法的一半，科学研究的作用就是使用实证数据来证明逻辑的有效性。"显而易见，数据不会自动地证明某种规律或逻辑的存在，规律往往隐藏在数据背后，需要人工借助方法和工具进行数据分析工作，数据分析就是从客观数据中找到客观规律和逻辑的中间过程，而这一过程本身的逻辑性、合理性必然影响到最终结果的发现和解释。纵观自然科学和社会科学各学科的科学研究活动、成果和文献报道，"量化研究方法作为学术训练的主体现象，普遍存在于各学科之中"。杜红亮和赵志耘回顾与展望了国内外软科学方法研究(2009)，指出了包括数理统计方法在内的多种方法的使用更为频繁，软科学方法研究将成为软科学研究重点，强调了各种方法的综合运用。

在量化研究的框架下，科学研究的基本元素是由数字构成的变量(Variable)，科学知识的基本单位则是描述变量与变量之间关系的假设。数据分析以数据为对象，抽样调查是取得数据的重要而且常用渠道之一，问卷中的测量项目，许多或多数是以选择题方式出现，形式上都是定性变量。由此可见，在社会和行为科学的科学研究中，对定性变量的数据分析成为重要领域

之一。

统计学(Statistics)方法作为传统的数据分析方法具有300多年历史,包括描述统计和推断统计,在自然科学、社会科学和人文科学领域中大量应用的是基于推断统计学理论的对抽样调查的数据分析。现如今,几乎所有领域都要用到统计学,许多文献都比较系统地讲述了经典统计学的基本理论,经典的推断统计学属于参数统计(Parametric Statistics),是用样本统计量推断总体参数,包括参数估计和假设检验。随着研究的深入,研究者发现参数推断对于解决许多数据分析问题具有局限。王星(2009)研究非参数统计时指出:"由于并不总是在参数的框架中找到答案,数据驱动的方法会带领数据分析的实践者突破传统框架。"于是在20世纪40~50年代诞生了非参数统计(Non-parametric Statistics),弥补了参数统计的许多不足。作为统计学的重要分支,非参数统计具有一些独特优势,特别是对于定性变量之间的关系分析,具有参数统计所不具备的作用,在许多条件下更有效。随着计算机的发展和数据的大量积累,在20世纪80~90年代逐渐兴起了数据挖掘(Data Mining,DM)技术,作为数据分析的重要方法。数据挖掘方法与统计学方法具有密切的关系,数据挖掘方法采用了许多统计学方法,但是仍然具有明显区别。统计学方法,不论是参数统计还是非参数统计,都是利用抽样调查所得数据,以样本状况推断总体状况;而数据挖掘方法的产生与数据库技术和数据库知识发现密切相关,所采用的数据是在数据库中积累的大量数据,所得结论就可以认为是总体状况,没有样本与总体的概念,因而没必要进行显著性检验。数据挖掘具有自己的独特方法,如关联规则方法,经过拓展完全可以适用于抽样调查中对定性变量的关联分析。王星(2009)在文献中将关联规则方法列入非参数统计中,这种融合无疑有利于数据分析的理论发展和实践应用。

在面对同样的变量、同样的客观数据情况下,参数统计、非参数统计和数据挖掘方法这三个数据分析的分支领域,所反映的思维角度和研究深度是不同的。统计方法,不论是参数统计还是非参数统计,首先都要对变量之间的关系做出假设,然后通过计算统计量、设立标准,最后做出接受或拒绝原假设的结论。既然是假设,不论是接受还是拒绝原假设都可能犯错误,即

两类错误之一，第一类错误（弃真错误）或第二类错误（取伪错误）。数据挖掘方法则没有事先的假设，其算法要求遍历给定范围内的所有数据和数据组合[武森，高学东，（德）M. 巴斯蒂安，2003]。不同的统计方法所犯两个错误的概率不同，不同的错误类型对于不同的研究内容和目标的风险不同，所以，在研究中不同的风险要求决定了所用方法不同。同理，不同的方法也反映不同的研究深度和角度，也就必然存在着一个根据研究深度和角度选择方法的问题。

数据分析的过程，实际上也是一个思维与表达的过程，1978 年诺贝尔经济学奖得主西蒙（Herbert A. Simon）在论述两者之间关系时认为：如果“唯有可表达的才是可思想的”这一假说站得住脚的话，那么“唯有可思想的才是可表达的”也一样正确。由此可见，思维活动，包括思考角度、对问题的抽象程度等，与表达方式是互动的关系，有什么样的思维，就应有与之相适应的表达；反之，表达方式应该正确反映研究者的思维。本书认为，思维与表达互动、互相适应这一思想，是指导数据分析的基本思想之一，如果违反这一思想，也就是表达与思维不一致，就会在研究过程或结论上产生矛盾。任何数据分析者在主观上都不愿意出现这种情况，但是之所以在客观上产生了表达与思维不一致的情况，其原因就在于方法的选用不当。在数据分析工作中，面对要解决的问题和同样的客观数据，可能具有不同的方法可供选择。在数据分析领域中同时研究统计学方法和数据挖掘方法的学者王星（2009）在其著作《非参数统计》中指出：“数据分析工作常常不是一个单纯方法的应用，而是对数据内部规律认识的从无到有的综合思考。”这一综合思考的结果，就是结合研究目的和数据特征选取正确的方法，方法的正确与否，涉及对研究对象研究的深度与广度不同，细致的程度不同，甚至是正确程度的不同。

随着计算机技术和数据分析软件的发展，人们面对大量数据，可以轻而易举地进行数据计算，关于数据分析软件的文献也比比皆是。但是这些文献往往从计算方法出发和软件功能出发，论述这些算法和软件模块的计算过程以及所能达到的目的。与此相反，从应用角度出发，根据不同研究目的比较、选择不同方法的文献比较稀少。

当研究者面对一整套调查数据时，为了理清分析的头绪、解决从何入手以及深入分析的路径问题，本书在以上提及或没提及的大师、诸多优秀学者的文献启发下，结合笔者的数据分析实践，思考了从研究的目标、广度、深度和不同角度等应用目标出发选择不同分析方法的问题，编纂出版以供研究者参考，希望起到抛砖引玉之作用。

1.2 问题的提出

1.2.1 分析的深度问题

分析的深度不够是许多问卷数据分析常见的问题，数据分析的任务就是全面描述数据特点、挖掘隐藏在数据背后的规律，从而揭示事物的本质状态、预测事物的发展趋势。在此将数据分析的深入程度由低到高分为四个层次，如图 1 -1 所示，由低到高逐层进行，同时每一层次都要结合专业理论和现实情况给予合理解释，事物的本质和规律是基于数据分析的合理解释。

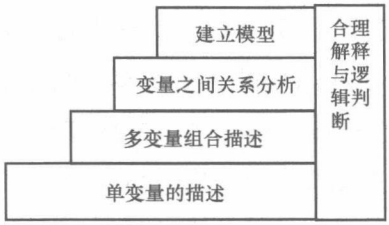


图 1 -1 数据分析的层次

上述四个层次的数据分析内容十分庞大繁杂，本书将范围限定在对问卷中选择题的分析处理上，一是缩小范围，二是旨在应用而不做过多的数学论证。

1.2.1.1 单变量的描述

单变量的描述是对选择题的各个选项进行频数和频率的统计，以表或图的形式表达，这是最基本的统计描述。

1.2.1.2 多变量组合描述

多变量组合描述包括两个层次:一是将问卷的问题或项目直接作为变量,称为项目变量,然后两两组合或多个之间组合,统计变量各个取值组合之间的数据分布,主要工具为列联表形式。二是对若干项目变量进行数学运算,构造出新的变量,称为构造变量,对构造变量再与其他变量进行组合描述。

1.2.1.3 变量之间关系分析

变量之间关系的分析主要分析两个变量之间的关系,包括:配合列联表的卡方检验、独立样本 t 检验、相关分析、回归分析、关联规则等,分析的对象同样包括项目变量和构造变量。后面单独有一章讲述该内容。

1.2.1.4 建立多元模型

对研究目标的关键问题和相应的各个影响因素建立一个模型,可以用公式、参数、表或图形表达,包括:对数模型、多元回归模型、关联模型、决策树模型等。

1.2.2 方法的选用问题

如果要分析得准确、深入和透彻,方法选用适当是基本要素,方法选用不当,可能造成许多问题,按照问题的严重程度大体上分为以下三个层次。

1.2.2.1 限于简单描述

在两个定性变量之间关系的分析中,列联表严格来说还不属于统计分析,还属于描述数据特征,而分析是指说明变量之间是否具有相关关系,相关的紧密程度如何,影响的方向如何,其方法可以用相关分析、回归分析等。

1.2.2.2 分析得不够深入

虽然有一定的分析,例如卡方检验、相关分析或回归分析,但深入程度不够,主要表现在缺少多元模型的建立,仅限于两个变量之间的关系分析。

1.2.2.3 分析结论与客观规律相悖

统计分析的目的是描述或发现数据总体上的规律,但是就少量数据个

体而言，完全有可能与总体数据特征不一致，这就出现了显著性检验问题。显著性检验通过设定临界点来区分变量差异的显著性，但是客观事物是连续变化的，有些边界情况难以生硬地划分其差异显著还是不显著，这就需要结合经验、常识和专业知识判断，如果经验告诉我们，两者差异应该是显著的，但是卡方检验却说明差异不显著，这就出现了悖论。在这种情况下，可以换一种方法，例如，两个比率之差的 t 检验，很可能就拒绝了无差异的原假设。所以说，在对抽样调查问卷的选择题进行分析时，特别是样本数量不足够大的情况下，对于数据分析的方法选择是否适当，是关系到研究过程与目标是否一致、研究结论是否正确的重要问题。这里所说的样本量，是指如果进行列联分析时，表中每格的样本量。有的研究者认为某项具有 1 000 个样本量的调查，应该不小了。如果进行二维列联分析，文化程度分为 6 类，职业分为 10 类，则该列联表有 60 个数据格，平均每格频数为 16，还是可以的，但是再加上性别这个变量，则会有 120 个数据格，平均每格频数为 8.3，事实上数据分布不可能均匀，所以有些数据格中的频数可能少于 5，这就不符合比较严格的分析条件了，如果再加上“年龄段”这个变量，样本量显然不够。上述只是导致分析结论与客观规律相悖的情况之一，还会有其他情况。

1.2.3 多视角综合分析问题

对某问题的分析，要从多视角入手，仅从单一视角往往难以发现事物的深层规律。多视角包括：对研究对象分类的多视角、对变量之间关联的多视角、对测量变量表象背后本质因素探索的多视角、对研究构念测量的多视角，等等。当然，多视角观察问题既包括问卷设计环节的考虑，也包括数据分析环节的考虑。如果认为在问卷设计完成后，观察问题的视角就已经确定了，这是片面的。在数据分析环节，完全可以根据现有测量变量进行运算组合，再构造出新的变量，也就是从新的角度考查所研究问题。

1.3 产生问题的原因

产生上述问题的原因，笔者主要归为三大类：第一是问卷设计缺陷；第二是缺少对数据分析方案的整体设计；第三是在分析中缺少根据所发现问

题构造变量再继续深入分析。

1.3.1 问卷设计缺陷

1.3.1.1 问卷设计要点

如果问卷设计有缺陷,则这种先天缺陷在以后的分析中难以弥补。问卷设计最重要的两个原则是准确和简约,显然这两个原则具有竞争性,问卷设计者要在这两个原则之间寻找平衡。一般来说,在简约方面,不要花费被调查者20分钟以上的时间;在准确方面,对于一个测量变量,常常要用4个以上的问题项目测量。谢家琳(2007)在"实地研究中的调查问卷"一文中提出了研究人员在设计问卷之前所应做的3点主要决策:

(1)问卷中将要调查哪些变量?要突出研究的重点。

(2)问卷中的变量之间是什么关系?一份典型问卷包括预测变量(解释变量)、结果变量和被调查者背景资料。

(3)问卷中所含的变量是什么样的结构?确定所要研究的变量由哪些观测维度组成,例如,工作满意度,其组成维度可以是:工作兴趣满意度、社会地位满意度、薪酬满意度、工作环境满意度、发展前景满意度、对上司满意度、对同事满意度等。

1.3.1.2 结构化设计问卷

在考虑问卷设计要点的过程中,可以采用结构化方法设计问卷,即按照自顶向下的原则,从研究目标开始,从所要调查问题类别分解到所要观测的变量,最终到测量项目。用框图的形式表达,如图1-2所示。

1.3.1.3 选项的形式及处理方式

(1)选项的形式。从选项的数量划分,选择题分为单项选择和多项选择;从测量的精度划分,在四大类测量尺度中,即定类变量(Nominal)、定序变量(Ordinal)、定距变量(Scale)和定比变量(Ratio),前两种属于定性变量,后两种属于定量变量。选择题的变量在表达形式上只是前三种类型,在本质上四种类型都有。在实际应用中,常常不区分定距变量和定比变量,在处理方法上也常都作为定量变量处理,而定序变量在许多情况下也常常按照定量变量处理。

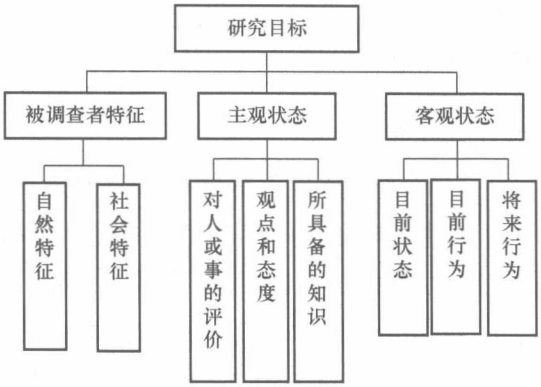


图 1-2　结构化问卷设计框架

从选项变量(测量变量)的关系划分:有些可以作为因变量(Dependent Variable),也称为依变量、结果变量或目标变量;有些变量可以作为自变量(Independent Variable),也称为解释变量、因素、分组变量等。因变量和自变量的关系是相对的,要根据专业知识、常识经验和定性逻辑判断。

(2)变量运算原则。尽管在实际应用中常常放宽条件,但是在概念上一定要明确变量运算的原则:定类变量只是类别不同,没有先后顺序或优劣之分,更不可以进行算术运算;定序变量只可比较先后顺序,不可以进行算术运算;定距变量只能做加减运算,不可以做乘除运算;只有定比变量才可以做各种算术运算。

在实际应用中,有些定序变量可以作为定比变量处理,有些则不可以,关键是要从变量的本质上考虑,是否违反变量处理的基本原则。对于在本质上属于定比变量,但是形式上作为定序变量的测量变量,在计算中仍然可以作为定比变量处理,否则不可以。例如:"收入"这个变量本来就是连续的定比变量,但是问卷的测量项目经常是将其作为封闭问题,例如:您的月收入是如下情况:

①<3 000　②≥3 000 ~ <6 000　③≥6 000 ~ <9 000　④≥9 000 ~ <12 000　⑤≥12 000

再例如,典型的李克特 5 及量表:您对本服务的满意度是:

①很不满意 ②部分不满意 ③一般 ④比较满意 ⑤很满意

由此形成的“满意度”这个变量,形式上是定序变量,但是其本质还是定比变量,因为人的感觉实际上是连续的,而且有正负之分,只不过限于表达方式,将其在形式上作为了定序变量。有些定序变量其本质就是定序变量,即只能按照定序变量处理,不可以进行算术运算。例如:您的行政级别是:①省部级 ②局级 ③处级 ④科级 ⑤科员。

1.3.1.4 多项选择题的优点

关于问卷设计有专门文献论述,在此提出一点需要注意,所设计的问题是单选题还是多选题,应该由问题的性质决定,尽量不要人为地规定。例如下面两个问题:

例题1:您的文化程度是:

①初中级以下 ②高中或中专 ③大专 ④本科 ⑤硕士 ⑥博士

例题2:您喜欢的金融理财方式有:

①银行存款 ②国债 ③股票 ④基金 ⑤投资型保险 ⑥黄金 ⑦期货 ⑧外汇买卖 ⑨代理理财产品 ⑩其他_____

在这两道例题中,题1是天然的单项选择题,而题2应该是多项选择题。但是笔者见到许多类似上述题2这样的问题,研究者人为地将其规定为单选题,将其变为:“您最喜欢的金融理财方式有(只选1项)”,这种将事实上的多选现象强硬地人为规定为单选题的做法让被调查者感到很不舒服,而且常常总会有人不听调查者的规定,仍然将其作为多选题,使得研究者将其作为无效问卷,而恰恰正是这些被调查者所提供的信息是有价值的。从研究方法上看,多项选择题相比单选题也有其优势,美国学者福勒(Floyd J. Fowler)在其专著《调查问卷的设计与评估》(Improving Survey Questions Design and Evaluation)(2010)提出了多项选择题比单项选择题具有的两个优点:①多项选择题可以更细致地,在更大范围测量所研究问题。②在一组多项目所具有的共同东西情况下,限制项目具有不良效果,而多项选择则可以更好地反映这些共同东西。

也许一些研究者认为多项选择题处理起来比较麻烦,同时许多文献没有论述如何处理,尽管有些方法很简单。针对此问题,本书设专门章节讨论