

万卷方法®

# 现代调查研究 方法和技术

陈心广 王培刚 著



全彩内文  
配套R代码



重庆大学出版社

武汉大学规划教材

# 现代调查研究方法和技巧

陈心广 王培刚 著



重庆大学出版社

### 图书在版编目(CIP)数据

现代调查研究方法和技术/陈心广,王培刚著--

重庆:重庆大学出版社,2023.10

(万卷方法)

ISBN 978-7-5689-4281-2

I. ①现… II. ①陈…②王…

III. ①社会调查-调查方法 IV. ①C915

中国国家版本馆 CIP 数据核字(2023)第 249269 号

### 现代调查研究方法和技术

XIANDAI DIAOCHA YANJIU FANGFA HE JISHU

陈心广 王培刚 著

策划编辑:林佳木

责任编辑:林佳木 版式设计:林佳木

责任校对:邹 忌 责任印制:张 策

\*

重庆大学出版社出版发行

出版人:陈晓阳

社址:重庆市沙坪坝区大学城西路 21 号

邮编:401331

电话:(023)88617190 88617185(中小学)

传真:(023)88617186 88617166

网址:<http://www.cqup.com.cn>

邮箱:[fxk@cqup.com.cn](mailto:fxk@cqup.com.cn) (营销中心)

全国新华书店经销

重庆升光电力印务有限公司印刷

开本:787mm×1092mm 1/16 印张:14.25 字数:316 千

2023 年 10 月第 1 版 2023 年 10 月第 1 次印刷

ISBN 978-7-5689-4281-2 定价:69.00 元

---

本书如有印刷、装订等质量问题,本社负责调换

版权所有,请勿擅自翻印和用本书

制作各类出版物及配套用书,违者必究

## 作者简介

**陈心广** 武汉大学特聘讲座教授。1977年进入武汉医学院(现华中科技大学同济医学院)卫生系学习,1992年在美国夏威夷大学东西方中心获得生物统计和流行病学博士学位。获得博士学位之后,先后在美国南加州大学、韦恩州立大学、佛罗里达大学从事教学和科研工作,担任助理教授、副教授和终身教授,博士生导师,同时也担任西安交通大学特聘讲座教授。作为主编,共同主编和编著过多部中英文专著和教材。在公共卫生、流行病学和全球健康方面展开了大量的科学研究,涉及分子、生物、心理、行为和社会等诸多领域,发表近300篇经同行评议的论文、专著章节和百科全书条目,美国流行病学院遴选院士。



**王培刚** 武汉大学公共卫生学院教授。南京大学与芝加哥大学联合培养博士,韦恩州立大学博士后。武汉大学公共卫生学院、政治与公共管理学院博士生导师,武汉大学人口与健康研究中心主任,《健康中国研究》集刊主编。长期从事人口与健康、健康中国等领域的研究。在《中国社会科学(内部文稿)》《中国行政管理》《光明日报》和 *Journal of Affective Disorders* 等中英文期刊、报纸发表论文数十篇,主持完成的研究报告有十余篇获得省部级以上领导肯定性批示或采纳,主持国家自然科学基金重大项目、国家自然科学基金重点项目等课题二十多项。以第一作者或独立作者身份出版了《社会变迁与中国居民生活质量》《多元统计分析与SAS实现》等中英文学术著作十多部(含译著),其中作为总主编在华中科技大学出版社出版“世界人口与健康经典译丛”(4本)。



## 前言

“调查研究是谋事之基、成事之道，没有调查就没有发言权，没有调查就没有决策权。”习近平总书记非常重视调查研究，并一以贯之身体力行、率先垂范，推动了全党大兴调查研究之风。调查研究有很多种类型，社会调查只是其中的一种。虽然名字是“社会调查”(social survey)，但是通过调查得到的数据却不仅能够用来研究解决社会问题，对分析解决很多其他方面的问题也有重要意义，包括经济发展、医疗卫生、人口发展等。而且许多社会调查运用的方法和技术，也可以用于其他领域的调查，比如环境资源调查、市场需求调查、医疗卫生服务调查等。

当今社会，大数据、人工智能和机器学习等数据科学快速发展，对调查数据的需求也与日俱增。通过调查研究获得数据，不仅速度快、数量大，而且质量高、可用性强。与一般的大数据不同，调查数据是为了解决问题而收集的，因此具有很强的针对性。除此之外，大多数大数据是没有固定结构的，而绝大多数通过调查得到的数据却是有结构的。比如，横断面调查的数据就是由 $n$ 个调查对象和 $p$ 个变量构成的数据矩阵。这样的数据，很方便直接进行处理分析。

学习社会调查技术和方法，首先必须明确社会调查主要用来解决哪些问题。对此，习近平总书记曾形象地指出：“坚持问题导向，增强问题意识，敢于正视问题，善于发现问题，既看‘高楼大厦’又看‘背阴胡同’，真正把情况摸清、把问题找准、把对策提实。”在总结个人经验和查阅文献的基础上，本书把调查研究的功能初步概括为五个方面。第一，用来描述问题的现状和把握事物的发展趋势；第二，用来发现和分析社会中存在的问题；第三，研究社会中各种事物之间的相互关系，尤其是因果联系；第四，用来预测、预报未来；第五，评价政策性干预项目的效果。

调查研究最吸引人的地方，就是通过调查得到的数据，可以让我们看到事物的真相。比如，只看好莱坞的电影，会以为西方国家到处都是美女帅哥，可是调查数据表明，美国

超重和肥胖的人占全国人口的比例在世界范围是最高的。又如,一个小城市的人去到大城市,被人偷了,就很容易得到大城市治安不好的印象,但是调查研究显示大城市往往比小城市安全。习近平总书记就指出了:“调查研究是做好各项工作的基本功。不了解真实情况,拍脑袋做决定,是做不好工作的。”

既然社会调查这么重要,那么,什么样的社会调查项目才能称得上是一个好项目呢?这就涉及对调查项目的评价问题。评价一个调查研究项目的好坏,常常需要从多方面考察。目前比较公认的评价标准有四个方面:1)重要性,好的社会调查项目,必须针对非常重要的问题;2)创新性,好的社会调查项目,一定要有创意、新意;3)可行性,即在人力、财力和技术方面是可以完成的;4)严谨性,即调查项目的立意、设计和实施方案是系统、周到、严密的。

一个好的社会调查研究之所以能够以一个小样本来反映总体的情况,达到去伪存真的目的,主要依赖于先进的方法和技术。这里指的是关于选题、调查问卷编辑、概率抽样、现场调查、数据处理和统计学分析等方面的技术和方法。随着信息科学日新月异的发展,新思想、新方法、新技术不断涌现,极大地丰富和提高了现代调查研究的能力。总体而言,我国在科学研究方面与发达国家的距离正在快速缩短,可是在社会调查方面,却还存在较大的差距。

引进社会调查的新技术、新方法,通过高等教育来学习是最重要的一条途径,而大学教育的重要环节,就是教材。当下,国内高校在社会调查方面的主流教材,其内容侧重在研究设计、抽样方法、调研方法等方面,很多学科的新进展、新的值得关注的重要内容还没有及时收录进来,比较典型的是答题心理学、卫星辅助抽样技术、单日重构调查技术、实况取样调查技术、敏感问题调查技术等。

本书作为一本教材,在编写的过程中,我们首先强调内容的**系统性**。一个社会调查项目,就是一个系统工程。从提出问题,到选题和课题构思,从现场调查设计、调查问卷编写、随机抽样,到现场调查的计划、管理与实施,再到最后的数据整理、数据库建立和统计学分析和推论,可谓环环相扣。在很多情况下,一项严谨的调查研究,还要考虑缺失值的处理、样本大小和统计学检验效率等问题。本书的编写,就是围绕这样一条主线展开的。

除了系统性,本书还特别注意内容的**新颖性**,在保证内容系统性的前提下,加入调查研究中的前沿内容,并让二者相辅相成。比如,在调查问卷编辑部分,介绍了调查研究中的答题心理学;在调查数据收集部分,介绍了调查数据产生和收集的理论模式和新的数

据收集技术和方法。这些理论、技术和方法,都能用于指导调查问卷的编写和修订、现场调查的计划和实施。熟悉掌握这些理论和方法,也能够提高研究人员自己的辨别能力和创新能力。

社会调查是一项实践性很强的工作,必须亲自上手才能学得更好。因此本书把可操作性放在重要的位置。在介绍各种方法和技术时,本书尽可能用实际例子,一步一步地演示如何实操完成。典型的例子包括手工和通过计算机进行随机抽样,用实际例子介绍如何编写调查题目、编排调查问卷,如何利用调查问卷里的信息给变量命名来建立原始数据库,通过程序演示数据处理、数据库的建立和管理、统计学分析和结果解释等。

除了调查设计和数据收集,统计软件的合理使用可以极大地提高研究的速度、效率和质量,因此用于数据处理和分析的计算机软件也是日新月异。一个典型的软件就是R。R是一款开源且可以免费使用的统计分析软件。该软件目前广泛用于大数据、人工智能、机器学习等数据科学领域。对R软件的安装和使用的介绍,成为本书中最具实用性和创新性的一部分。结合书中的内容,我们还用R编写了一系列演示程序,来帮助读者学习数据库的建立、数据处理和统计学分析过程。

本书的内容,主要源于两个作者在国内外的教学和调查研究实践,同时参考了下列教材和专著:*Quantitative Epidemiology* (Chen, Springer Nature, 2021)、*Statistical Methods for Global Health and Epidemiology* (Chen & Chen, Springer Nature, 2020)、*Survey Methodology* (Groves et al., Wiley Inter-science, 2014)、《现代社会调查方法》(风笑天,华中科技大学出版社,2014)、《全球健康研究方法》(郝元涛、陈心广,人民卫生出版社,2018)、《多元统计分析与SAS实现》(王培刚,武汉大学出版社,2020)。

编写教材是一项非常严肃的工作。我们编写这部教材的动机,是通过系统介绍当前比较前沿的理论、方法和技术,来促进我国社会调查研究的发展。在本书的出版、校对过程中,马弋茗博士等一众同学做出了突出贡献,在此表示感谢。由于本书是我们的第一次尝试,书中瑕疵在所难免,希望能够得到谅解。如果发现问题,请批评指正,以便后期修订时进行参考。

陈心广 王培刚

2023年4月25日



本书练习使用数  
据和R软件代码  
可扫码下载

# 目录

<b>第一章 调查研究概论</b>	<b>1</b>
1.1 调查研究的意义	2
1.2 调查研究与大数据	6
1.3 调查研究构思与选题	7
1.4 选题的评价标准	9
1.5 描述性调查研究	12
1.6 影响因素和因果关系研究	14
1.7 干预效果评价研究	16
1.8 本书的目的和内容安排	17
<b>第二章 现场调查设计</b>	<b>20</b>
2.1 明确调查对象和选择调查类型	20
2.2 横断面调查设计	23
2.3 多波横断面调查设计	25
2.4 病例对照调查设计	28
2.5 纵向调查设计	30
2.6 随机对照设计	34
<b>第三章 调查研究常用抽样技术和方法</b>	<b>38</b>
3.1 研究总体、研究对象和样本	38
3.2 非概率抽样方法	41
3.3 简单随机抽样	42
3.4 随机整群抽样	46
3.5 两阶段和多阶段整群抽样	49
3.6 系统抽样	49
3.7 分层整群抽样	51
3.8 卫星辅助概率抽样	52
3.9 滚雪球抽样和调查对象驱动抽样	54

<b>第四章 调查问卷的编辑方法和技术</b>	<b>60</b>
4.1 答题心理学	60
4.2 调查问卷的基本结构	62
4.3 调查问卷主体结构 and 单个调查模块	64
4.4 调查题的编写技术	66
4.5 提高数据质量的调查题设计编写技术	68
4.6 调查问卷主体内容的设计	73
4.7 调查问卷内容的编排指南	77
4.8 调查问卷初稿、预实验和订正	78
<b>第五章 现场数据收集方法与技术</b>	<b>84</b>
5.1 现场调查的操作模式	84
5.2 调查数据的产生模式	86
5.3 现场调查所需技术设备和方法	88
5.4 焦点小组研究和结构性访谈	89
5.5 群体(集体)调查	92
5.6 入户调查	94
5.7 邮寄调查	94
5.8 日记调查方法	95
5.9 单日重构调查方法	95
5.10 实况取样调查方法	97
5.11 现场调查的管理与质量控制	98
<b>第六章 调查数据处理和数据库建立的方法和技术</b>	<b>102</b>
6.1 调查数据整理和录入	102
6.2 电脑或在线调查数据的整理	107
6.3 创建数据库	108
6.4 运用R进行数据库管理	109
6.5 数据处理	116
6.6 编制数据库编码手册	119
<b>第七章 基本统计学分析</b>	<b>124</b>
7.1 连续型变量的统计学描述	125

7.2	离散型变量的统计学描述	130
7.3	连续型变量的两组和多组比较分析	133
7.4	离散型变量的两组和多组比较分析	137
<b>第八章</b>	<b>线性相关和回归分析</b>	<b>143</b>
8.1	线性相关分析	144
8.2	简单线性回归分析	149
8.3	多元线性回归分析	152
8.4	logistic回归分析	155
<b>第九章</b>	<b>调查数据缺失的评价与插补</b>	<b>162</b>
9.1	导致数据缺失的主要原因	162
9.2	数据缺失的类别和评价	164
9.3	数据缺失的统计学机制	168
9.4	用平均数替代缺失值	170
9.5	缺失值的多重插补技术	170
9.6	敏感性分析	175
<b>第十章</b>	<b>样本大小和统计检验效率分析</b>	<b>179</b>
10.1	样本大小和统计检验效率分析简介	179
10.2	比较样本率或构成比的样本大小和统计检验效率	182
10.3	比较两个样本均数的样本大小和统计检验效率	185
10.4	比较多个样本均数的样本大小和统计检验效率	188
10.5	相关分析时的样本大小和统计检验效率	190
10.6	多元回归分析的样本大小和统计检验效率	192
10.7	确定样本大小和统计检验效率的经验总结	194
10.8	补救(充)性统计检验效率分析	196
<b>附 录</b>		<b>200</b>
	附录1:武汉市农民工社会资本与行为健康调查问卷(节选)	200
	附录2:“2019中国社会状况综合调查”调查问卷摘选	208
	附录3:主要术语表	210

# 第一章 调查研究概论

## 没有调查就没有发言权

毛泽东同志的著名论断“没有调查就没有发言权”和“调查就是解决问题”，是对“调查研究”的目的和作用的精辟概括。习近平总书记也十分重视调查研究，他说：“调查研究是谋事之基、成事之道，没有调查就没有发言权，没有调查就没有决策权。”

以人为对象的科学知识和技术成果，很大一部分都建立在调查研究所得数据的基础之上(Chen et al., 2018)。这些成果表现在众多领域，包括社会学、心理学、行为科学、经济学、预防医学、临床医学、药学、公共卫生政策和医疗机构管理等。作为研究对象的人，既是一种客观的存在，又是一种主观的存在。因此，但凡涉及人的研究，一般都不可能像数学、物理、化学和机械工程等基础学科那样，把研究对象直接分解成很小的、相对独立的部分，进行有控制的实验性研究。认识个人的行为、心理、态度以及社会方方面面的现象，了解事物的现状和发展趋势，掌握其影响因素，制定相应的干预措施和评价干预结果，都需要开展调查研究(Groves et al., 2004; 风笑天, 2014; Chen, 2021)。在数据科学(data sciences)，包括大数据、机器学习和人工智能)方兴未艾的今天，调查研究更具有特殊的价值。调查研究为数据科学提供大量高质量的数据，是大数据的重要数据来源。同时，用于调查研究的思想、理论、方法和技术，也可以帮助和推动大数据和数据科学的发展(Japac et al., 2015)。

调查研究(survey study)作为一种常识，我们每个人都很熟悉。对于调查研究的了解，大多数人局限于早期革命和战争年代的社会调查。早期的社会调查，一般是定性的。这类调查所用的通常是所谓的“解剖麻雀”的方法。具体做法是：发现问题，寻找典型案例，从典型案例中寻找关键人物，开展面对面访谈(face-to-face interview)。访谈完毕后，根据个人记忆和笔录，组织不同范围的集体讨论，形成一些初步结论。再通过核心人员的思辨、推理和直觉等心理过程得出结论。然而，基于科学技术的现代调查研究，大多以定量研究为主。在确定了研究目标和调查对象之后，运用抽样技术选择有代表性的样本，采用专业的方法和技术从调查对象那里收集数据，对数据进行加工整理后，再用定量分析方法进行分析，从数据里提取所需要的信息，最后通过这些提取的信息，来回答研究一开始提出的问题(Groves et al., 2004; Chen, 2021)。

## 1.1 调查研究的意义

调查研究的重要性不言而喻,因为没有调查就没有发言权,仅仅凭个人的判断始终无法消除偏见。比如,生活在贫民窟的人,每天一开门满街看到的都是穷人,就觉得社会穷。相反,生活在富人区的人,一开门见到的就是豪车、豪宅,便觉得国家很富。即使不带个人偏见,仅凭少数人获得的信息,也无法准确和全面地反映客观情况。传统的社会调查也有局限,因为数据仅仅来源于几个典型代表,无法保证调查结果能够反映全局的情况(Pronin, 2007)。因此必须借助于科学的调查研究,才能够更准确、客观地了解情况,最终发现问题、解决问题。再比如,经常看好莱坞的电影,你可能会以为在发达国家到处都是帅哥美女,然而事实却恰恰相反。大量调查数据结果表明,发达国家人群超重或肥胖的比例高达68%(Wang et al., 2020),即每10个人中约有7个超重或肥胖。

为什么基于科学的调查研究结果比简单的观察更好呢?首先,调查研究获得的数据来源于通过专门技术从总体中抽取出的一群有代表性的个体。其次,调查研究得到的数据是以科学原理为依据,以系统框架为要求,按照标准程序来收集的。最后,调查结果是应用专业的统计学分析方法计算和检验得到的,真正做到了“用数据说话”。通过调查收集到数据之后,采用现代技术和方法进行统计学分析,通过分析从数据中提炼出反映客观事物的信息,了解和把握总体的情况,发现存在的问题,找到事物之间内在的关系,采取针对性的措施来解决现实生活中的问题。基于此,科学的调查研究是通过调查数据,而不是个人的观察和主观判断来得出结论。

如图1-1所示,调查研究的重要性可以从多个方面来理解。第一个方面是去伪存真,这是调查研究的基本性质;第二个是把握事物现状,这是调查研究最基本的应用;第三个是了解事物的发展趋势,指导未来决策;第四个是发现存在的问题,这是调查研究把握基本情况的前提,通过对比来完成;第五个是研究事物间的因果关系或影响因素;第六个是评价重大干预措施的效果。下面就这些方面分别进行讨论。

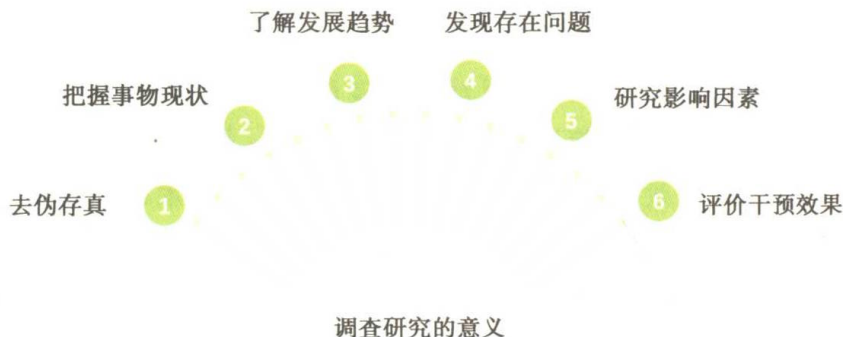


图 1-1 把握调查研究的六大意义

### 去伪存真

从本节一开始所举的例子不难看出,人的天性之一就是习惯于把自己日常生活中所经历的事实,当成普遍结论。然而,这种“事实”往往并不具有普适性,可能与真相有很大

的偏差。这样的现象不胜枚举。比如,一个人在自己的微信群里看到有很多人购买某一种东西,就认为全社会的人都在购买;养老院的工作人员,可能会认为社会上到处都是老年人,可是到2020年底我国65岁以上人口占总人口的比例仅13.5%;到医院里走一趟,可能会认为到处都是病人;而到健身房一看,全部都是健康人。一个病人到某三甲医院治好了病,不仅感谢主治医生,而且到处宣传这个医院的服务水平和服务效率如何好;而另一个病人听了这个病人的话,到相同的医院找了相同的专家却没有治好。因向上帝祈祷而免于意外死亡(如暴风雨、车祸)的人相信上帝,却无法知道那些遭遇相同的意外而没有活过来的人是否也同样向上帝祈祷过(Dean, 2005)。

如果每个人都坚持自己的看法,社会就无法正常运转,人们就无法正常生活。科学的调查研究,就是为了消除这种偏见而提供的一个解决方案。科学的调查研究要求,先从研究总体(population)中随机地选择一个小于总体但又能够代表全体的样本(sample),再用科学的方法从调查样本中每一个调查对象那里收集数据。最后通过对数据的统计学分析和检验,来推断真实总体的情况。比如,通过大范围随机抽样调查,可以客观准确地判断一个国家和社会的贫富状况,如个人收入、家庭人口、幸福感、获得感等。通过健康调查,可以知道一个社会中健康的人和生病的人的比例,拥有医疗保险的比例,看病的人次数等。通过对医院进行调查和对比分析,可以把不同的医院按照服务质量排名,同时指出不同等级医院的强弱项。通过严格的调查,也可以研究证明信仰宗教和向上帝祈祷是否与健康和疾病有关系(Goldman, 2004)。

### 把握事物的现状

调查研究常用来帮助人们把握事物的现状。几乎所有的社会问题,都可以开展调查研究来了解现状,诸如生老病死、教育、婚姻、家庭、社会、政治、经济和文化生活等。典型的社会问题包括每年某地区有多少人出生? 多少人死亡? 多少人生病? 多少人结婚或离婚? 多少人失业? 收入水平和贫富差别怎么样? 社会稳定和治安情况如何? 开展调查研究,就能够用科学的数据来回答这些问题,这也是为什么调查研究会在实际工作得到广泛的应用。

为了把握事物现状而进行的调查研究,一般都属于描述性的(descriptive)调查。这是很重要的一种调查类型,本章后面的1.5节还要专门介绍。在技术方面,描述性调查一般用横断面设计(cross-sectional design)。关于横断面调查设计,在第二章会有详细介绍。横断面调查必须用随机样本(random sample),才能够反映总体情况。随机抽样技术(random sampling techniques)会在第三章进行详细介绍。调查数据的统计学分析,将在第七章和第八章专门进行介绍。

### 了解发展趋势

调查研究的重要性还体现在它能够帮助把握事物的发展趋势,预测未来的发展方向。只要有两次及以上的横断面调查数据,就可以用来描述事物的发展趋势。如果有更多次的横断面数据,就可以建立预测模型,估计事物的未来发展趋势。从技术的角度来

看,横断面调查简单易行;从经济的角度看,横断面调查成本低;从实际操作方面来看,横断面调查速度快、成效显著。

世界上很多国家都积累了多年的重复横断面调查数据。最典型的例子就是美国的“综合社会调查”(General Social Survey, GSS)。中国也于2003年正式创建了“中国综合社会调查”(Chinese General Social Survey, CGSS),由中国人民大学中国调查与数据中心负责,这是中国最早的全国性、综合性、连续性学术调查项目。

此外,中国政府相关部门也牵头组织了一些大型多波次重复横截面调查,如中国流动人口动态监测调查(China Migrants Dynamic Survey, CMDS)。这项调查起始于2009年国家卫生健康委员会组织开展的一年一度的大型全国抽样调查,数据采集一直持续到2018年。调查内容包括多方面的信息,如个人和家庭情况、城乡迁移流动经历、社会福利保障、健康生活方式、生命质量、医疗卫生服务、婚育子女、社会融入等。CMDS为把握我国流动人口的基本情况、工作就业、日常生活、社会活动、健康状况及流动人口管理和政策制定提供了不可或缺的科学数据。利用这些数据也培养了多名硕士和博士研究生。

由原卫生部组织开展的全国卫生服务调查也属于反复多次横截面调查的典型例子,第一次卫生服务调查开始于1993年,每五年开展一次,延续至2018年已完成第六次,是我国政府掌握城乡居民健康状况、医疗卫生机构和人员、卫生服务价格和利用程度、医疗保健费用及负担等信息的重要途径,也是中国卫生调查制度的重要组成部分。其调查结果对政府制定卫生政策和卫生发展规划、有效调控卫生服务供求关系、提高卫生管理水平、促进我国卫生改革与发展产生了重要影响,也为推动中国的卫生事业改革发展提供了重要的数据支持和科学指导。

### 发现存在的问题

除了把握现状、历史趋势和预测未来,调查研究的重要性还体现在它是发现重大问题的科学手段。众所周知,发现问题有时比解决问题更重要。通过调查研究来发现问题,往往是与把握现状和了解历史趋势同时进行的。比如,社会发展和居民健康的横断面调查发现,随着经济的快速发展,困扰西方社会的超重和肥胖,也成为威胁中国人民健康的一个非常重要的问题。而且根据反复多波横断面调查的结果进行预测分析发现,超重和肥胖的情况在我国未来十几年里,还会进一步恶化(Wang et al., 2020)。同样地,通过横断面出生率调查,科研人员发现中国的少子化问题严重,每年新出生的人数还不足以替代每年死亡的人数。如果不及时采取有效的措施,少子化会极大地影响国家的长期发展。

一般而言,调查研究是发现一个地区公众最关心问题的非常有效的手段。“得众则得国,失众则失国。”比如,通过调查研究,可以发现低收入的省、市、县、乡、镇、村和低收入的人群。找到后,就可以采取针对性的扶贫措施,来减少或消除贫富差距,达到共同富裕的目标。通过调查研究,可以发现农民工的城市融合、就业、子女上学、住房和医疗保险等许多问题。发现问题后,就可以通过一些政策措施,帮助农民工融入城市,参与城市的

建设和发展。还可以通过调查研究界定常见疾病,用于指导医疗卫生资源的合理配置和有效使用。此外,调查研究还可以找出国民健康水平发展较为落后的地方和人群。据此,政府层面可以指导医疗和公共卫生决策向落后地区和人群倾斜,加强疾病预防、治疗和康养活动,维护并促进公众健康,消除健康不平等现象,“不让任何一个地区和一个人掉队”,最终实现健康中国的总体目标。

### 研究影响因素

研究影响因素和认识事物间的因果关系,对所有学科的研究人员都有极大的吸引力(Chen, 2021; 风笑天, 2014)。这是因为,只有找到影响因素才能够真正把握问题发生的原因和机理,才可以对症下药地采取干预措施。与实验室研究不同的是,调查研究往往着重从人群整体层面,从宏观角度来探索、分析和了解影响问题的主要因素,探索这些因素与问题之间的关系。典型的例子包括:工业化、城市化与肥胖、糖尿病、心脏病和癌症之间的关系;农民工进城与基础建设和经济发展之间的关系;互联网与生活质量之间的关系;社会资本与健康之间的关系;等等。后文1.6节还要对这方面的内容进行深入讨论。研究影响因素和因果联系需要特别的调查设计,比如病例对照设计(case-control design)和纵向追踪设计(longitudinal follow-up design)等,本书第二章将有专门介绍。

通过调查研究得到的因果关系有两大用处。第一,可以用来指导更深入的实验性研究。比如,通过调查发现社会资本与生活质量有正相关关系。为了回答为什么社会资本会提高生活质量的问题,就可进一步研究社会资本影响生活质量的机制,包括社会支持、资源共享、信息共享、个人价值认同、心理韧性等。再比如,通过大范围调查发现工业化水平与人群肥胖的比率呈现明显的正相关。根据这一发现,可以开展更深入的研究,来认识工业化通过什么途径导致肥胖发生率的增加。第二,可以依据调查研究结果开发出具有针对性的干预措施,来解决实际问题。比如,通过调查研究发现,缺乏体力活动是影响健康和生产力的一个重要因素。为了保护生产力和提高健康水平,就可以针对不同的人群,开发出不同的能够促进体力活动的干预措施,以便更广泛地实行(Chen, 2021; Pearl, 2000)。

### 评价干预效果

调查研究是评价干预措施(intervention measure)在人群中效果的科学手段。但凡涉及大众的、与国计民生相关的干预措施,都可以通过调查研究来评估(assessment)。典型的例子包括新发疾病防控措施(如新冠病毒的防控措施)、新的医疗保险制度、新的课外辅助教育政策、新的招生政策、新的生育政策、新的农民工户口管理政策、新的环境治理政策……政策相继出台和实施,但其是否有效果?如果有,效果有多大?回答这些问题的主要方法就是调查研究。鉴于调查研究对干预项目评估的重要性,本章第1.7节还要进行更深入的讨论。

在传统的社会治理过程中,虽然贯彻执行了很多政策和措施,但鲜有运用现代调查研究方法来评价效果的。通常的做法是找几个典型案例,召开总结大会,集体进行

学习或列举一些事实了事。这种做法,完全不能满足现代社会治理的要求。必须用现代调查研究技术和方法进行定量评价,确定干预政策和措施是否有效,达到预期设定目标。在评价的同时还有可能发现新的问题,及时针对性采取补救措施。因此,通过现代调查研究而不是传统的总结大会来评估政策性干预措施,应该成为现代社会治理的一个重要标志。

## 1.2 调查研究与大数据

进入本世纪以来,伴随着信息科学和技术的快速发展,大数据、机器学习、人工智能和数据科学等新兴学科方兴未艾。有的人甚至认为传统的调查研究和统计学分析已经过时(Dunson & DB, 2018)。事实上,随着科学的进步和社会的发展,人们发现新兴的信息科学技术与传统的现场调查和统计学分析并不冲突,而是能形成互补(Japiec et al., 2015)。调查研究是大数据的一个重要数据来源,同时大数据研究也要借用调查研究的理论和方法。下面从两个方面展开讨论。

### 调查研究是重要的数据来源

在大数据时代,任何事物都可能成为数据,包括微信的聊天记录、电话的通话记录、自媒体发表的信息记录、个人每天的行踪等。当然,大数据中也少不了通过长期各种各样的现场调查所积累的数据。因此,调查研究是当代大数据、数据科学、机器学习和人工智能等新兴学科的一个重要的、高价值的和不可替代的数据来源。

之所以说调查数据是重要的,是因为一个大型的调查研究,往往都是围绕一个或几个非常重要的问题,在很大的范围内开展的。比如,一个国家的人口和就业情况,居住和社会环境条件,儿童的教育,居民健康水平和常见病,医疗卫生机构发展现状和存在的问题,医疗保险类别和覆盖情况,等等。每一个待研究的问题,往往是从很多可能的问题里,通过专家评审、行政审批等复杂的过程确定下来的。与此同时,调查研究一般都是由受过专业培训的科研人员来完成的。因此,正规的调查研究所提供的数据,往往具有很高的科学价值。

通过调查研究获得的数据,往往以标准化的格式存放(见本书第六章)。因此,通过调查研究得到的数据,一般都可以直接进行统计学分析。需要注意的是,目前所称的“大数据”,绝大多数都不能直接进行分析。因为此类数据均为对现实生活的直接记录,其目的不是为研究而收集。此类数据更像流水账,虽具有客观性,但无固定结构。在进行定量分析之前,必须花时间和精力进行处理,把无结构(unstructured)的数据转变为有结构(structured)的待分析数据。

### 现场调查研究与大数据相辅相成

无论是大数据、机器学习、人工智能还是数据科学,不管其名字多么响亮、方法多么高级,其目的都是一样的,即掌握情况、预测未来、理解因果关系、评价干预措施、促进社