

抽样调查

Sampling Survey

YINGYONG TONGJI YU XINXI CONGSHU

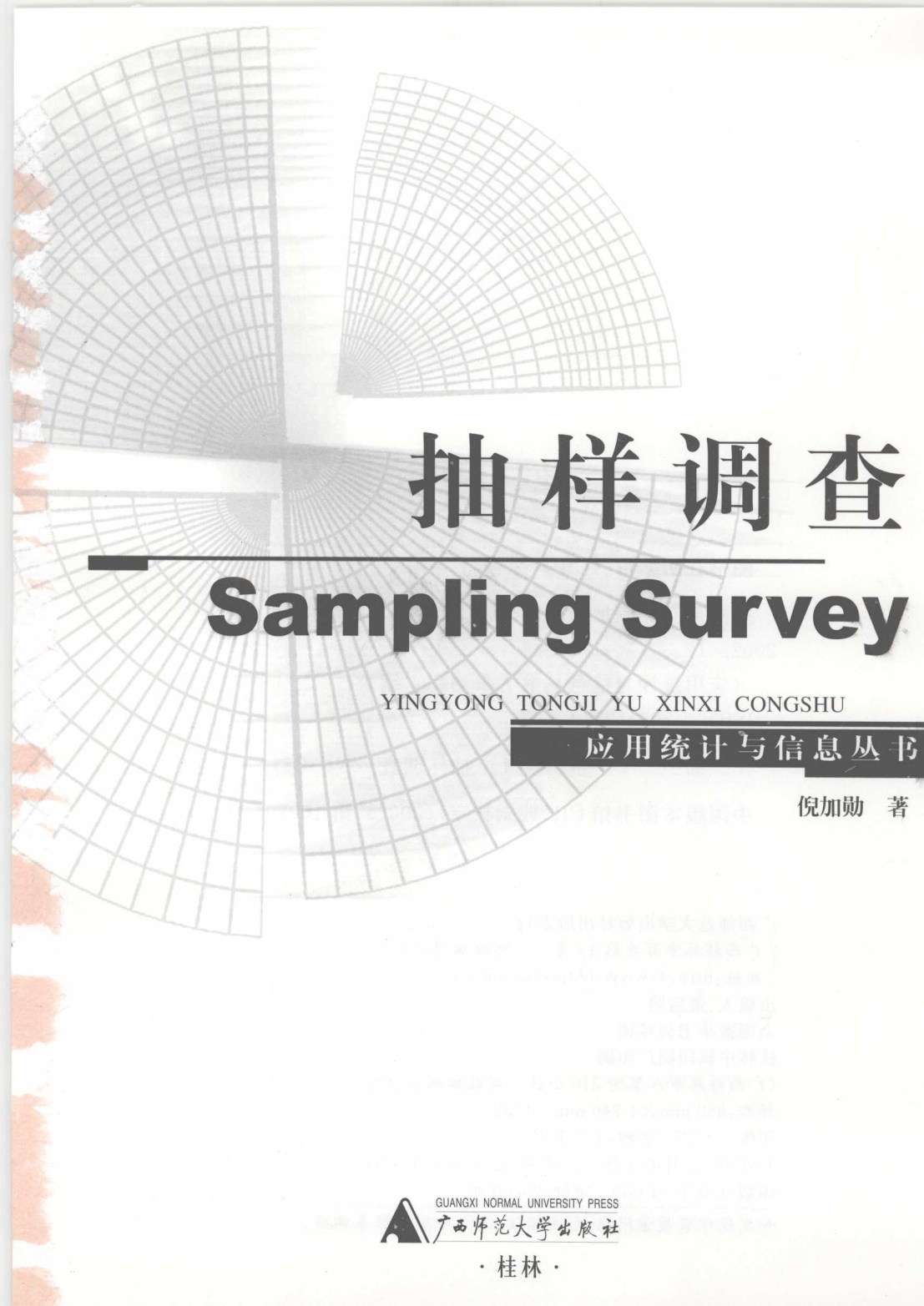
应用统计与信息丛书

倪加勋 著



GUANGXI NORMAL UNIVERSITY PRESS

广西师范大学出版社



抽样调查

Sampling Survey

YINGYONG TONGJI YU XINXI CONGSHU

应用统计与信息丛书

倪加勋 著



GUANGXI NORMAL UNIVERSITY PRESS

广西师范大学出版社

· 桂林 ·

图书在版编目 (CIP) 数据

抽样调查 / 倪加勋著. — 桂林: 广西师范大学出版社,
2002. 12

(应用统计与信息丛书 / 张尧庭主编)

ISBN 7-5633-3772-5

I. 抽… II. 倪… III. 抽样调查 IV. C811

中国版本图书馆 CIP 数据核字 (2002) 第 100294 号

广西师范大学出版社出版发行

{ 广西桂林市育才路 15 号 邮政编码: 541004 }
{ 网址: <http://www.bbtpress.com.cn> }

出版人: 萧启明

全国新华书店经销

桂林中核印刷厂印刷

(广西桂林市八里街 310 小区 邮政编码: 541213)

开本: 890 mm × 1 240 mm 1/32

印张: 12.25 字数: 353 千字

2002 年 12 月第 1 版 2002 年 12 月第 1 次印刷

印数: 0 001 ~ 1 000 定价: 20.00 元

如发现印装质量问题, 影响阅读, 请与印刷厂联系调换。

出 版 说 明

即将到来的 21 世纪是一个信息的时代,掌握信息、了解信息、处理信息是企业、政府、个人 必须面对的问题,信息处理中统计方法是十分重要的,因此我们出版这一套《应用统计与信息丛书》,来推动这一方面的发展.

我们聘请了国内统计与信息方面的著名专家,撰写一系列的著作,结合国内实际,介绍国外 的动向,发挥自己的专长,就一个个主题作深入浅出的讨论,既能使广大的读者受益,又有 相当的学术水平,把我国的应用统计与信息处理推向一个新的高峰.

希望这套丛书的出版能受到读者的欢迎.

广西师范大学出版社

1998 年 8 月

前 言

抽样调查是统计学的一个分支,它是获取统计数据的一种重要手段,广泛应用于各个领域,并为世界各国所普遍采用.我国的政府统计工作已把抽样调查作为主要的调查方式列入了统计法.目前我国各大学统计系普遍开设了这一课程.本书是作者多年来在中国人民大学给本科生和研究生讲课的基础上形成的,在编写过程中也参考了近年来国内外出版的有关抽样调查的专著和教材吸取了一些较好的例子和习题,以及一些简明的证明(见本书附录中的主要参考书目),还注意了以下几点:

1. 兼顾抽样调查的理论和实际应用,但重点放在应用.在介绍各种抽样调查方法的同时,尽量说明其使用的条件,实施的方法,讨论怎样使用才能提高抽样效率,等等.对于公式,除了一些比较繁杂、占用篇幅过多的证明予以简略外也尽量给出证明.

2. 便于自学和理解.在介绍内容时尽量举一些例子来加以说明.本书的例子有两类,一类是介绍每一种方法以后如何应用的例子;另一类是对一些初学时比较难理解的概念和方法,如群内相关系数、PPS和 π PS抽样方法、比率估计中无偏的估计方法等,用一些简单的数字、直观的方法来举例说明.

3. 考虑了读者的广泛性,尽量适应各种读者的需要.本书每章后都有一些思考和练习题供读者用来检验是否真正理解和掌握了该章的主要内容,本书既可以作为教材使用,也可以作为参考资料或自学用书.对于不需数学证明的读者,可以把证明部分跳过,阅读方法和例子.但本书后面的附录一又可供想进一步了解数学证明的读者提供了一些预备知识.

4. 增添了一些新的内容.如缺乏抽样框情况下对总体容量的估计.本书介绍了捕捉——再捕捉的抽样方法和条样的密度估计等方法.

因为在目前新的形势下如欲估计盗版书籍的数量,现有上网的网民、网站等数量,这些方法可以借鉴应用。

本书是在张尧庭教授的推荐和广西师大出版社余鑫晖总编的一再鼓励下着手编写的,由于作者水平有限,错误在所难免,希望读者批评指正。

倪加勋
于中国人民大学
2001年3月

目 录

第一章 绪 论	(1)
§ 1.1 抽样调查是获取统计数据的一种重要手段	(1)
1.1.1 社会发展与统计数据	(1)
1.1.2 数据的种类	(2)
1.1.3 统计获取数据的要求	(3)
1.1.4 数据的获取方法	(5)
§ 1.2 抽样调查的特点和作用	(6)
1.2.1 什么是抽样调查	(6)
1.2.2 抽样调查的作用	(7)
1.2.3 抽样调查的发展历史和目前的应用	(9)
§ 1.3 抽样调查的基本原理和常用概念	(14)
1.3.1 抽样调查的基本原理	(14)
1.3.2 抽样调查的一些常用概念	(18)
思考与练习(共8题)	(26)
第二章 简单随机抽样	(28)
§ 2.1 引 言	(28)
2.1.1 什么是简单随机抽样	(28)
2.1.2 简单随机抽样的实施方法	(30)
2.1.3 简单随机抽样在抽样理论中的地位 and 作用	(31)
§ 2.2 总体均值和总体总量的估计	(32)
2.2.1 简单随机抽样中应用的符号	(32)
2.2.2 总体均值的估计	(33)
2.2.3 总体总量的估计	(36)
§ 2.3 总体比例的估计	(39)
§ 2.4 样本容量的确定	(42)

2.4.1	估计总体均值 $\bar{Y}$ 和总体总量 Y 时样本容量 n 的确定	(43)
2.4.2	估计总体比例 P 时样本容量的确定	(45)
2.4.3	要求控制相对误差时样本容量的确定	(46)
2.4.4	对稀有元素的寻找及估计比例时样本容量的确定	(48)
2.4.5	多指标情况下样本容量的确定	(50)
2.4.6	对总体的各子总体要求分别估计并达到预定精度时 样本容量的确定	(50)
2.4.7	总体方差或总体差异系数的估计	(51)
§ 2.5	子总体的估计	(53)
2.5.1	子总体均值的估计	(53)
2.5.2	子总体总量的估计	(55)
	思考与练习(共8题)	(58)
第三章	分层抽样	(61)
§ 3.1	引言	(61)
3.1.1	什么是分层抽样	(61)
3.1.2	分层抽样的作用	(61)
3.1.3	分层抽样的实施	(62)
3.1.4	分层抽样的一个直观例子	(63)
3.1.5	分层抽样的符号	(64)
§ 3.2	总体均值、总量和比例的估计	(65)
3.2.1	总体均值的估计	(65)
3.2.2	总体总量的估计	(67)
3.2.3	总体比例的估计	(69)
3.2.4	总体中具有某种特征单位数的估计	(71)
§ 3.3	样本量在各层的分配	(72)
3.3.1	样本分配对精度与费用的影响	(72)
3.3.2	等比例分配	(73)
3.3.3	奈曼分配	(75)
3.3.4	最优分配	(78)

3.3.5 近似分配	(80)
§ 3.4 样本容量的确定	(80)
3.4.1 确定样本容量的一般公式	(80)
3.4.2 等比例分配的样本容量公式	(83)
3.4.3 奈曼分配的样本容量公式	(84)
3.4.4 最优分配的样本容量公式	(85)
§ 3.5 分层抽样的效率分析	(88)
3.5.1 分层抽样与简单随机抽样比较	(88)
3.5.2 用分层样本数据计算分层抽样的设计效果($Deff$) ..	(89)
3.5.3 奈曼分配与按比例分配的比较	(92)
3.5.4 层权误差对效率的影响	(93)
§ 3.6 一些特殊的分层技术	(94)
3.6.1 事后分层	(94)
3.6.2 目录抽样	(96)
3.6.3 根据频数资料的分层	(99)
§ 3.7 多指标分层抽样中的一些问题	(101)
3.7.1 有控制抽选	(101)
3.7.2 多指标情形下样本量在各层的分配	(105)
3.7.3 多指标分层抽样样本量的确定	(106)
思考与练习(共 14 题)	(106)
第四章 比率估计和回归估计	(110)
§ 4.1 引言	(110)
4.1.1 比率估计及其应用	(110)
4.1.2 使用比率估计法的目的	(113)
4.1.3 比率估计中使用的符号	(113)
4.1.4 回归估计	(114)
§ 4.2 比率估计	(114)
4.2.1 总体比率 R 的估计	(114)
4.2.2 总体均值 $\bar{Y}$ 和总体总量 Y 的比率估计	(118)
4.2.3 利用比率估计提高抽样效率的条件	(122)
4.2.4 样本容量的确定	(124)

6.1.3	整群抽样中群的划分	(193)
6.1.4	整群抽样的符号	(194)
§ 6.2	群的大小相等的整群抽样	(195)
6.2.1	总体均值和总体总量的估计	(195)
6.2.2	总体比例的估计	(198)
6.2.3	选择群的大小	(199)
§ 6.3	群的大小不等的整群抽样	(204)
6.3.1	比率估计方法	(205)
6.3.2	无偏估计的方法	(210)
6.3.3	不等概率抽样	(212)
6.3.4	三种方法的比较	(215)
§ 6.4	整群抽样的效率	(217)
6.4.1	群内相关系数 ρ_c	(217)
6.4.2	整群抽样的效率	(221)
6.4.3	利用样本数据计算整群抽样的效率	(222)
§ 6.5	样本量的确定	(224)
6.5.1	等概率抽样比率估计的样本量	(224)
6.5.2	采用无偏估计方法的样本量	(226)
6.5.3	PPS 抽群的样本量	(228)
6.5.4	根据设计效果确定样本量	(228)
思考与练习(共 14 题)		(229)
第七章 系统抽样		(234)
§ 7.1	引言	(234)
7.1.1	什么是系统抽样	(234)
7.1.2	应用系统抽样的主要原因	(235)
7.1.3	系统抽样的抽样方法	(235)
§ 7.2	系统抽样的特点	(240)
7.2.1	系统抽样与其他抽样方法的关系	(240)
7.2.2	系统抽样方差的不同表示方式	(241)
§ 7.3	系统抽样的估计	(247)
7.3.1	无关标志排列的系统抽样	(247)

7.3.2 按有关标志排列的系统抽样	(248)
7.3.3 有周期性波动下的系统抽样	(251)
§ 7.4 其他系统抽样	(253)
7.4.1 不等概率系统抽样	(253)
7.4.2 分层系统抽样	(255)
7.4.3 二维系统抽样	(257)
思考与练习(共 10 题)	(258)
第八章 二重抽样	(262)
§ 8.1 引言	(262)
8.1.1 什么是二重抽样	(262)
8.1.2 二重抽样的作用	(262)
§ 8.2 为分层而进行的二重抽样	(264)
8.2.1 二重分层抽样的抽选方法	(264)
8.2.2 二重分层抽样的估计	(265)
§ 8.3 为比估计而进行的二重抽样	(268)
8.3.1 二重抽样比率估计的样本抽选	(268)
8.3.2 二重抽样的比率估计	(269)
§ 8.4 为回归估计而进行的二重抽样	(271)
8.4.1 二重抽样回归估计的样本抽选	(271)
8.4.2 二重抽样的回归估计	(272)
§ 8.5 为 PPS 抽样而进行的二重抽样	(276)
8.5.1 二重 PPS 抽样的样本抽选	(276)
8.5.2 二重 PPS 抽样的估计	(276)
§ 8.6 二重抽样样本量的最优分配	(279)
8.6.1 二重分层抽样的最优分配	(279)
8.6.2 二重抽样比估计时的最优分配	(282)
8.6.3 二重抽样回归估计时的最优分配	(284)
思考与练习(共 7 题)	(285)
第九章 多阶段抽样	(287)
§ 9.1 引言	(287)
9.1.1 什么是多阶段抽样	(287)

9.1.2	多阶段抽样的作用	(288)
9.1.3	二阶段抽样中的符号	(289)
§ 9.2	初级单位大小相同的二阶段抽样	(290)
9.2.1	总体均值与总体总量的估计	(290)
9.2.2	总体比例的估计	(294)
§ 9.3	初级单位大小不等的二阶段抽样	(296)
9.3.1	二阶段等概率抽选,无偏估计	(296)
9.3.2	二阶段等概率抽选,比率估计	(299)
9.3.3	按初级单位大小的 <i>PPS</i> 抽样	(301)
9.3.4	估计比例的二阶段抽样	(305)
§ 9.4	二阶段抽样的扩展	(309)
9.4.1	分层二阶段抽样	(309)
9.4.2	三阶段抽样	(311)
§ 9.5	二阶段抽样中样本的最优分配	(314)
9.5.1	二阶段不分层的情况	(314)
9.5.2	分层二阶段的样本分配	(316)
§ 9.6	二阶段抽样的效率	(317)
9.6.1	二阶段抽样与简单随机抽样比较	(317)
9.6.2	二阶段抽样与分层抽样比较	(318)
9.6.3	二阶段抽样与整群抽样比较	(318)
	思考与练习(共 12 题)	(319)
第十章	一些其他的抽样方法	(323)
§ 10.1	总体单位数(容量)的估计	(323)
10.1.1	直接抽样法	(323)
10.1.2	逆抽样方法	(326)
10.1.3	采用条样(Quadrat Samples)估计总体密度和数量	(329)
§ 10.2	敏感性问题的抽样方法	(332)
10.2.1	什么是敏感性问题	(332)
10.2.2	沃纳(Warner)随机化回答模型	(333)
10.2.3	西蒙斯(Simmons)随机化回答模型	(335)

10.2.4	随机化回答模型的进一步应用	(337)
10.2.5	使用随机化回答方法应注意的问题	(339)
§ 10.3	自加权抽样设计	(339)
10.3.1	什么是自加权抽样设计	(339)
10.3.2	单阶段分层抽样的自加权抽样设计	(341)
10.3.3	单阶段分层 PPS 抽样的自加权抽样设计	(341)
10.3.4	不分层的二阶段抽样的自加权抽样设计	(342)
10.3.5	分层二阶段抽样的自加权抽样设计	(343)
§ 10.4	关于连续性的抽样	(346)
10.4.1	连续性抽样中的一些问题	(346)
10.4.2	样本轮换的最优比例	(346)
10.4.3	估计两个时期之差的最优轮换比例	(349)
10.4.4	估计两个时期之和的最优轮换比例	(350)
§ 10.5	关于无回答的处理	(350)
10.5.1	抽样中无回答的影响	(350)
10.5.2	利用二重抽样处理无回答	(352)
10.5.3	不进行再次访问的调整	(354)
§ 10.6	每个家庭抽选一个成年人的方法	(358)
10.6.1	Kish 方法	(358)
10.6.2	Deming 方法	(360)
10.6.3	二维随机表方法	(362)
	思考与练习(共 12 题)	(362)
附录一	本书的预备知识	(365)
附录二	随机数字表	(376)
主要参考书目		(377)

第一章 绪论

§ 1.1 抽样调查是获取统计数据的一种重要手段

1.1.1 社会发展与统计数据

社会发展离不开统计数据.统计作为一种计数的活动由来已久.我国远在夏商时期就已有人口和土地的记载,而且在治国中具有重要作用.春秋时期齐桓公任管仲为相,使齐国大治,在反映管仲思想的重要著作《管子》一书中就曾指出:“不明于计数而欲举大事犹无舟楫而欲经于水险也.”*意思是说,不掌握数据而想进行大的事业,就好像要过河而没有船和桨,是很危险的.在战国时期的商鞅,也提出要以数据作为制定政策的依据,他说要使国家富强,必须掌握十三项数字,“境内仓口(府)之数,壮男壮女之数,老弱之数,官士之数,以言取食者之数,利民之数,马牛刍藁之数”,并强调说:“欲强国,不知国十三数,地虽利,民虽众,国愈弱至削.”**再看看西方国家的情况,英国的威廉大帝为了征兵和征税,于公元 1086 年曾下令编制全国土地调查清册,其内容还包括当时的人口、牲畜、农具等.由于调查过程中官吏横加威逼,对于农民犹如教会所说的末日审判,故该资料有“末日审判书”(Domesday Book)之称.该资料现存放于伦敦档案馆,成为研究 11 世纪末英国社会状况的珍贵史料.

随着社会的发展,就更加需要对各种社会经济现象作定量研究.尤其是社会化大生产的出现、资本主义的产生,促使自然经济瓦解、人口向城市集中,在资本主义发展较早的英国,生活出现两极分化,并多次

* 见《管子》的《七法》

** 见《商君书》的《去强》篇

出现瘟疫,引起人们对人口问题的关心.英国的格朗特(J. Grunt)于1662年出版了《对死亡表的自然观察和政治观察》一书,研究了人口统计中的一些数量规律.与此同时,国家之间争夺殖民地的竞争加强,英国的威廉·配第(W·Petty)写了著名的《政治算术》一书,用数字和计量的方法把英国经济和其他国家进行比较.他们这些通过数字来研究社会经济问题,使统计学开始萌芽.在另一方面,自然科学的发展,如农业的田间试验、生物的遗传变异、天文学的气象预测、化学中的化合与分解等,都需要数量分析.这些不确定现象的研究,随着数学分支概率论的发展,成了统计学诞生的催化剂.统计学的产生,反过来又促进了数据的搜集和处理功能,它不仅为治理国家和科学研究提供了有力的手段,而且还成为企业管理的重要工具,如用于产品的质量管理、企业的预测和决策等领域.

现代科学技术的发展,尤其是以计算机为核心的信息处理技术的发展和推广,不仅使传统的农业、工业和其他产业的生产效率不断提高,信息本身也逐渐形成一个独立的部门,越来越多的人转向从事信息的处理、传递和贮存的有关工作.人们意识到充分的信息资源和有效的信息的处理技术是正确决策的必要条件,它会产生巨大的财富和新的变革.信息已渗透到各行各业,人们称这种变化为步入信息时代,也有人称之为数字化时代.统计是处理数据获取信息的重要手段,统计资料是信息的一个重要组成部分,由此可以看出,社会愈发展,对统计数据的需求也愈强烈.

1.1.2 数据的种类

客观世界十分复杂,任何事物都是质和量的统一.由于数据的特点不同,就决定了获取数据方法的多样性.从数据的性质及如何去搜集的角度,数据可以分为两大类:调查数据和试验数据.

(1) 调查数据

调查数据是指客观现象在自然的状态下发生,通过调查或观察来加以搜集和整理.如社会经济现象中,在一定时期或某一时点某种现象的状况,包括它的规模、水平和现象之间的相互关系.如一个国家的人口、工业产量、农作物的播种面积等,它们是研究社会经济必不可少的

数据,而且这些现象随着时间的变化而变化,因而需要经常性的定期的调查.另外有一些是有关部门为了管理的需要或是研究部门、学术机构对某一现象的研究而进行的专题调查,此外还有市场调查、民意测验等所获得的都是调查数据.

(2) 试验数据

其特点是在人为的控制条件下来搜集数据,通常与自然科学的研究相联系.其特点是试验之前数据尚未产生,因而需要通过事先的设计,在一定的条件下进行试验,并将试验的过程和结果加以记录和整理,这样得到的数据,称为试验数据.试验数据与调查数据的不同还表现在调查数据通常是无法重复试验,而试验数据则可以在控制的条件下重复进行.如化肥的增产效果,农药的杀虫效率,某种化学变化合适的温度、压力和时间等.用试验收集数据的方法,有时也可应用于社会经济现象,如产品质量的比较检验、商品包装对销售量的影响等.试验数据往往和试验的条件有关,数值也随着条件的变化而变化.由于试验是可以重复进行的,因而试验的次数是无限的,不可能进行全面观察而只能用抽样的方法.

由于上述两种数据的性质不同和研究的目的不同,因此就要分别采用调查和试验两种不同的方法.它们各有一些特殊的方法,在统计学中已形成统计调查和试验设计两个分支.本书的内容主要叙述如何有效地进行统计调查.

1.1.3 统计获取数据的要求

统计工作不仅要提供数据,而且要从数据中揭示出客观现象及其规律性,从而为决策提供依据.因此如同工厂生产产品一样,要求提供优质的产品,需要有好的原材料,统计数据就是统计工作和统计分析的原材料.俗话说:“巧妇难为无米之炊.”工厂有好的原材料才能加工出好的产品,而统计需要有好的数据才能正确描绘客观事物和揭示客观现象之间的联系,从而为正确决策提供依据.那么怎样才算好的数据呢?统计获取数据有以下要求:

(1) 科学性

这里的科学性有两重含义:首先统计指标的设计必须符合客观实