目 录

第1章　绪论

1.1　统计与统计学

1.1.1　“统计”的含义

在日常生活中“统计”有着多种含义。例如，企业管理人员要掌握生产销售情况和利润额；在生产车间要统计产品生产量；在高考录取中要统计考生的总分；球类比赛时解说员要统计竞赛双方的进攻次数和成功率；开会时主持人要统计一下出席会议的人数，等等。人们也常常从报刊杂志、电视新闻中获悉我国的经济增长速度、消费者价格指数、固定资产投资规模等经济数据资料。那么到底什么是“统计”呢？

请思考：下列资料中“统计”一词的含义是什么？

(1)张宏是学统计的；

(2)他已搞了几十年统计了；

(3)据统计，今年一季度物价指数出现负增长；

(4)请找统计登记一下；

(5)请统计一下今天的销售量。

那么，把统计作为一种专业用语，其含义到底是什么？目前，关于统计一词的含义比较趋于一致的解释为：一是统计工作，二是统计数据，三是统计学。

统计工作，即统计实践，是对社会经济现象客观存在的现实数量方面进行搜集、整理和分析预测等活动的总称。如上所列举，要统计进攻次数、统计生产量、统计交易额等，就是统计工作。一个完整的统计工作过程一般包括统计设计、统计调查、统计整理、统计分析等环节。统计工作是统计一词最基本的含义，是人们对客观事物的数量表现、数量关系和数量变化进行描述和分析的一种计量活动。如银行的计划统计科，每月编制项目报表，这个过程就是统计工作。又如：我国进行人口普查时要经过方案设计、入户登记、数据汇总、分析总结和资料公布等一系列过程，这些都是统计工作。在我国，各级政府机构基本上都有统计部门，如统计局，它的职能主要是从事统计数据的搜集、整理和分析工作。

统计资料(统计信息)是统计工作过程中所取得的各项数字资料和与之相关的其他实际资料的总称。经济增长速度、价格指数等，就是统计资料。它和前面讲的统计工作是紧密相联的，是统计工作的结果，因而也是很早就有的。根据历史记载，我国夏禹时代就开始有人口统计数字。春秋时期，《商君书》中指出：强国知十三数，这十三数包括粮食储备、人口及其各项分类数、农业生产资料以及自然资源等，不过当时还没有明确叫做统计资料罢了。随着社会的发展，需要的统计数

字也就越来越多，现在只要打开报纸就可以看到各种各样的统计数字。国家统计局每年出统计年鉴，反映国家的经济、文化教育以及科技发展等情况，这些都是在这个意义上的统计。如：

(1)我国国土面积960万平方公里，其中山地约320万平方公里，高原约250万平方公里，平原约115万平方公里，丘陵约95万平方公里。

(2)2003年我国全年全部工业增加值53612亿元，比上年增长12.6%，其中规模以上工业企业(即国有工业企业及年产品销售收入500万元以上的非国有工业企业)增加值增长17.0%，工业产品销售率98.1%，比上年提高0.1个百分点。这些由文字和数字共同组成的数字化的信息就是统计资料，是统计提供数据信息的基本表现形式，是统计工作的直接成果。统计资料包括原始资料和整理后的资料即次级资料。例如企业各车间的统计台账、人口普查时初次登记的资料就是原始资料，而统计公报、调查分析报告等现实和历史资料就是次级资料。统计资料的表现形式有统计表、统计图、统计分析报告、统计公报和统计年鉴等。

除了上面所讲的两个含义之外，“统计”一词还有另外的含义，即作为一门科学的“统计学”，它是本书将要探讨的主要内容，作为一门科学的“统计学”，它的出现要比统计工作和统计资料晚得多。

1.1.2 统计学的产生与发展

统计起源很早，是随着社会生产的发展和国家管理的需要而逐步产生和发展起来的，距今已有四五千年的历史。而统计学或统计理论则是在长期统计实践活动基础上形成和发展起来的，距今只有300多年的历史。回顾一下统计的渊源及其发展过程，对于我们了解统计学的研究对象和性质，学习统计学的理论和方法，提高我们的统计实践和理论水平，都是十分必要的。

1. 统计实践的产生和发展

在原始社会，人类最初的一般计数活动蕴藏着统计萌芽。在奴隶社会，统治阶级为了对内统治和对外战争，需要征兵征税，开始了人口、土地和财产的统计。例如，公元前3050年，埃及建造金字塔，为征集建筑费，就有对全国的人口与财产调查；罗马皇帝恺撒·奥古斯都曾下过一道命令，要全世界向他纳税，于是要每个人都向就近的收税人登记；英国的威廉大帝下令测量英国的土地，其目的是为了征税和征兵役；我国在夏朝就有了关于人口和土地的数字记载：夏朝时分中国为九州，人口约1355万人，土地约2438万顷；春秋时期齐桓公任用管仲为相使齐国大治，在反映管仲思想的重要著作《管子》一书中就有这样的论述：“不明于计数，而欲举大事，犹无舟楫而欲经于水险也”，这就是说不善于利用计数而进行宏伟事业，犹如没有船和桨而想渡过激流险滩一样。

由于生产力水平所限，奴隶社会的统计只属于初级阶段。到了封建社会，统计有了一定的发展，封建君主和精明的政治家日益意识到统计对于治国强邦的重要性，统计范围有所扩大。但由于封建经济的封闭割据和保守性，统计活动的范围受到限制，统计方法也很不完善。到了资本主义社会，随着社会生产力的迅速发展和社会分工的愈益精细，统计得到了很大的发展，统计的应用领域越来越广泛，不仅仅只局限于经济管理领域，在军事、医学、生物、物理、化学等领域中也大量地运用统计方法。而电子计算机技术的应用为统计活动的现代化进程提供了重要手段。正是在这样的历史背景下，统计学应运而生。从17世纪中末期开始，经过300余年的发展，形成了今天的统计学。

2. 统计学的产生和发展

在统计学作为一门科学逐渐形成的过程当中，由于历史和社会的原因，形成了很多学派。

(1)国势学派。国势学派也称记述学派，是伴随德意志的兴盛而产生的。其代表人物是康令(H. Conring，1606—1681)和阿亨瓦尔(G. Achenwall，1719—1772)。国势学以叙述国家显著事项和国家政策关系为内容，并给出了统计学这一名词，但国势学主要用文字表述，缺乏数字内容，和现代统计学比较起来，有一些实不副名，即没有太多的实际内容符合统计学的名称。

(2)政治算术学派。政治算术学派产生于 17 世纪的英国，代表人物是威廉·配第(W. Petty，1623—1687)和约翰·格朗特(John Graunt，1620—1674)。威廉·配第著有《政治算术》一书，在书中提出“不用比较级、最高级进行思辨或议论而是用数字来表达自己想说的问题，借以考察在自然中有可见的根据的原因”。该书用数量分析的方法对比了英国、法国和荷兰三国的“财富和力量”，以批驳当时英国国内的悲观论调。在书中他用数字来表述，用数字、重量和尺度来计量，并配以直观的图表，为现代统计学提供了一个完美的开端。马克思称威廉·配第为“政治经济学之父，在某种程度上也可以说是统计学的创始人”。

约翰·格朗特对英国伦敦人口的出生率和死亡率分别进行统计，并编制出世界上第一张“死亡率”统计表。他们的工作和使用方法，在现代统计学中，也是重要的，但遗憾的是，他们没有使用“统计学”一词，后人戏称“名不符实”，即没有用适当的名称称呼其实际的统计工作的内容。

国势学派和政治算术学派同时开始于 17 世纪中叶，直到 19 世纪末，从不同的国度、地域做了统计学的奠基工作。从统计学的内容上看，描述统计学的基本内容和方法已经给出；从时间上划分，这两个半世纪的统计学研究，又可称为“古典统计学”。并且随着微积分的完善、概率论的发展，为统计学注入新的内容，共同为推动统计学的形成打下了基础。

(3)数理统计学派。产生于 19 世纪中叶，创始人是比利时的天文学家、数学家和统计学家凯特勒(1796－1874)，其著作有《统计学的研究》、《关于概率论的书信》等。他是当时统计学界的中心人物，担任过比利时中央统计局局长，主持过第一次国际统计会议(1853 年)，他最先将概率论应用于人口、人体测量和犯罪等问题的研究，完成了统计学和概率论的结合。从此，统计学开始进入更为丰富发展的新阶段。许多学者从各个角度研究统计学，不断增加新内容，相继提出和发展了相关和回归理论、t 分布以及抽样理论等，使数理统计学很快发展成为一门比较系统、完善的学科。国际统计学界称凯特勒为“近代统计学之父”，就在于他发现了大量现象的统计规律和开创性地应用了许多统计方法，促使统计学向新的境界发展。由于这一学派主要在英美等国发展起来，故又称英美数理统计学派。

数理统计学派在理论上混淆了自然现象和社会现象之间的本质区别，过分夸大了概率论的作用，认为统计学就是数理统计学，是现代数学的一个分支，是通用于研究自然现象和社会现象的方法体系，否认社会经济统计学的存在，因而又导致了与社会经济统计学派的长期争论。

(4)社会经济统计学派。这一学派于 19 世纪后半叶兴起于德国，即原来政治算术意义下的统计学。但由于它在理论上比政治算术学派更加完善，在时间上比数理统计学派提前成熟，因而它很快占领了“市场”，对国际统计学界影响较大，流传较广。其主要代表人是恩格尔(1821－1896)和稍后的梅尔(1841－1925)。他们主张统计学是研究社会现象的社会科学。这一学派融会了记述学派和政治算术学派的观点，并把政府统计和社会调查融合起来，进而形成社会经济统计学。

数理统计学派与社会经济统计学派共存并争论至今已有 100 多年，目前，虽然数理统计学派在国际统计学界占据着优势，但二者已出现了融合的趋势。

统计发展史表明，统计学是从设置指标研究社会经济现象的数量开始的。随着社会的发展与实践的需要，统计学家对统计方法的不断丰富和完善，统计学也不断发展和演变。从当前世界各国统计研究状况来看，统计学已不仅为研究社会经济现象的数量方面，也为研究自然技术现象的数量方面提供各种统计方法；它既研究确定现象的数量方面，又研究随机现象的数量方面。从统计学的发展趋势来看，它的作用与功能已从描述事物现状、反映事物规律，向抽样推断、预测未来变化方向发展。它已从一门实质性的社会性学科，发展成为方法论性质的综合性学科。

1.1.3 统计学的定义

古往今来的统计学者对统计学给予了不同的定义。根据美国统计学家 David Freedman 等著的《统计学》(魏宗舒等译，中国统计出版社，1997 年版)中的定义：统计学是对令人困惑的问题作出数字设想的艺术。

根据《蓝登书屋大字典》(the random house college dictionary)，统计学是“对用数字表示的事实或数据进行收集、整理、分类、分析以及解释的科学”。简而言之，统计学就是统计数据的科学。

统计学的定义众多，比较有代表性的是如下定义：

定义 1.1 统计学：统计学是收集、整理、显示和分析统计数据的科学，其目的是探索数据内在的数量规律性。

统计学的定义告诉我们，统计学是关于数据的科学，其内容包括数据收集、整理、显示和分析等。数据收集也就是取得统计数据，数据整理和显示就是将数据用图表等形式展示出来，数据分析则是通过统计方法分析数据，并对分析的结果进行说明。

把统计学称为艺术显然有些夸张，但这一说法的目的正在于提示统计工作者，应当创造性地提出和解决统计问题，不应局限于某些条条框框去理解统计这门科学。

案例：在一个水库中养着许多鱼，管理人员希望了解鱼的大致数量。这就是一个实践中的统计学问题。由于鱼是不听从指挥，会在各处自由游动的，因此，在进行统计时，必须创造性地提出解决方案。一种解决方法是先从水库的不同位置一共捕上来 1000 条鱼，在每条鱼的尾部作上一个标记，应当保证标记不会影响鱼的自由游动。然后，将鱼全部放回水库。几天后，从水库中再捕上来 2000 条鱼，检查其中尾巴上有标记的鱼的数量。假定在第二次捕上来的 2000 条鱼中，有 20 条尾巴上做了标记，则可以推断，水库中鱼的总数大致为

$$1000/(20/2000)=10 \text{ 万条。}$$

上述这个案例在实践中是经常见到的，对于一个统计工作者来说，作出一个 10 万条鱼的估计是不够的，他还应当对这一估计的精度作出判断。但这种搜集统计数据的方法，更多的是一种艺术，是很难从书本上学到的。

统计学是一门艺术。它提供一种归纳推理的方法，推理就是一种艺术。统计的应用方面是十分复杂的，只有将统计理解为一种艺术，创造性地提出新的方法去解决新的问题，才是真正地掌握了统计的精髓。既然是归纳推理，就不能保证结论百分之百正确，就不能没有争议。怎样让别人看懂并理解统计理论，就要看统计表达这些结论的技巧和艺术性了。

统计学是一门科学。它提供一套方法和技术，这些方法和技术并不是一成不变的，使用者在给定的情况下必须根据所掌握的专门知识选择使用这些方法，而且，如果需要还要进行必要的修正。统计方法是通用的数据分析方法，这些方法不是为某个特定的问题领域而构造的。

统计学是一种工艺。如同工业生产过程中的质量控制程序一样，统计方法是在为保证产品达到所希望的质量和保持其稳定性的管理系统中建立起来的。统计方法也能用于控制、减少和

考察不确定性。

统计学是科学、工艺和艺术这三者的组合。

统计学是研究大量社会现象(主要是经济现象)的总体数量方面的方法论的科学。这里所指的方法论包括指导统计活动的原理原则、统计过程所应用的核算和分析的方法以及组织方法。人们通过对社会现象中各种数量关系的研究来认识社会现象发展的规律性。值得注意的是,统计学在研究社会现象时,首先从定性研究开始,即在搜集原始统计资料(统计调查)之前,就要根据所要研究的对象的性质和研究任务、目的,确定调查对象的范围,规定分析这个对象的统计指标、指标体系和分组方法,这种定性工作是下一步定量分析的必要准备。在定量分析的基础上再达到认识社会经济现象的本质、特征或规律,这就是质－量－质的统计研究过程和方法。

1.2　统计学与统计数据

1.2.1　统计学的研究对象及特点

统计学的研究对象是指统计研究所要认识的客体。只有明确了研究对象,才可能根据它的性质特点指出相应的研究方法,达到认识对象客体规律性的目的。由统计学的发展史可知,统计学是从研究社会经济现象的数量开始的,随着统计方法的不断完善,统计学得以不断发展。因此,统计学的研究对象为大量现象的数量方面。

统计学的特点可以归纳为以下几个方面:

1. 数量性

社会经济统计学最基本的研究特点就是以数字为语言,用数字说话。具体地说,是用规模、水平、速度、结构和比例关系等,去描述和分析社会经济现象的数量表现、数量关系和数量变化,揭示事物的本质,反映事物发展的规律,推测事物发展的前景。

但应注意,统计学研究现象的数量方面,不同于数学上研究的纯数量,它不是抽象的数量,它是以现象质的规定性为基础的,是带有一定具体内容的数量。因为任何事物都是质和量的辩证统一,没有质也就没有量。

例如:要了解哈尔滨市重工业产值,首先要明确什么是重工业。所谓重工业是为国民经济各部门提供技术装备、动力和原材料的工业,包括采掘工业、原材料工业和制造工业。然后要确定重工业产值的含义和统计口径以及哈尔滨市哪些企业属于重工业企业,这些都是质的规定。在此基础上,还要解决怎样搜集、整理和汇总重工业产值资料,最后才能得到哈尔滨市重工业产值的具体数值。

2. 总体性

总体性又称大量性或综合性。统计研究的着眼点是大量社会经济现象总体,而不是少量或个别现象,它是通过对个别事物大量观察,占有丰富材料,加以分析综合,来反映现象总体的数量特征,揭示现象的本质和规律性。例如,2004 年全年居民消费价格总水平比上年上涨 3.9%,这个数量反映的是 550 多种消费商品及服务项目价格总的平均上涨水平,而不是指哪一种具体消费商品或服务项目的价格上涨水平。而要对这 550 多种消费商品及服务项目的价格上涨情况进行调查,就必须先对每一种个别消费商品及服务项目的价格情况进行调查,然后进行汇总

综合，从而达到对550多种消费商品及服务项目价格的总体认识。

统计研究并不排除从个别现象入手，但统计研究个体是为了综合个体而认识总体，是手段而不是目的，其最终目的是要认识总体。例如，2000年11月1日进行的第五次全国人口普查，逐一登记了全国大陆31个省、自治区、直辖市（不包括香港特别行政区、澳门特别行政区、台湾省）的每个人的性别、年龄等特征，但人口普查的目的并不是要了解关于某个人的特征，而是为了通过对全国人口情况进行汇总计算，得出关于我国人口总体的特征资料，从而达到对全国人口现象总体的认识。汇总后结果显示，祖国大陆31个省、自治区、直辖市（不包括福建省的金门、马祖等岛屿，下同）和现役军人的人口共126583万人。同第四次全国人口普查1990年7月1日0时的113368万人相比，十年零四个月共增加了13215万人，增长11.66%。平均每年增加1279万人，年平均增长率为1.07%。同1990年第四次全国人口普查相比，0－14岁人口的比重下降了4.80个百分点，65岁及以上人口的比重上升了1.39个百分点。从总体着眼，从个体入手，体现了统计工作中总体和个体之间的辩证关系。

3. 社会性

社会经济统计学通过研究大量社会经济现象总体的数量方面，来认识人类社会活动的条件、过程和结果，反映物质资料的占有关系、分配关系、交换关系以及其他的社会关系。其定量研究是以定性分析为前提的，而定性使其在客观上就有了社会关系的内涵。社会经济现象与自然科学技术问题是不同的，对于同一社会经济现象，站在不同的立场，持有不同的观点，运用不同的方法，可以得出差别较大的结论。这些都体现出社会经济统计活动的社会性。

4. 变异性

变异性又称差异性。统计学研究同类现象总体的数量特征，它的前提是总体各单位的特征表现存在着差异，而且这些差异并不是由某种固定的原因事先给定的。例如一个地区的居民人口有多有少，居民的文化程度有高有低，住户的生活消费水平有升有降，等等，正是各单位之间这种差异的存在，才需要研究地区的人口总数、居民文化结构、住户平均生活消费水平等统计指标。如果各单位不存在这些差异，也就无需进行统计，如果各单位之间的差异是按已知条件事先可以推定的，也就无需进行统计调查研究。

5. 具体性

统计研究的总体数量是一个有具体时间、具体地点、具体条件限定的数量。如利润额800万元，单独看来，它只是一个毫无意义的抽象数量。如果说2004年12月某企业利润额800万元，这就是统计中所说的具体数量了。可见具体性就是指在时间、地点、条件三方面有着明确的规定性。

统计工作虽然是研究具体的数量，但为了进行复杂的定量分析，还需要借助抽象的数学模型和数理统计方法，遵循一定的数学规则。以抽象方法为手段，以具体数量为目的，体现了统计研究中具体和抽象的辩证关系。

1.2.2 统计学的学科构成

统计学是一个复杂的体系，不同的学术机构对于这门学科的构成进行了不同的阐述。

美国数学学会出版的《数学评论》中对统计学的分类：

A. 基础；B. 充分性和信息；C. 决策理论；D. 抽样理论和抽样调查；E. 分布理论；F. 参数推

断；G. 非参数推断；H. 多元分析；I. 线性推断；J. 试验设计；K. 序贯分析；L. 随机过程推断；M. 工程统计学；N. 应用；O. 统计表。

从非统计专业的学生学习的角度来看，统计学可以分为四个大的组成部分：

1. 调查与实验设计

调查与实验设计涉及统计中获得原始数据的各种方法。调查是在社会经济统计中获得原始数据的主要手段。随着市场经济的发展，调查在经济活动中所起的作用越来越大，企业的经营，政府的决策，都离不开来自调查的第一手数据。

在科学研究过程中，获得统计数据的手段还包括实验方法。实验是在研究对象进行一定控制的情况下获得数据的方法。

2. 描述统计

描述统计包括整理、显示和分析数据的一系列方法。调查或者实验中所获得的有关事物整体的原始资料，往往是零乱和不系统的，需要经过一系列的统计处理，才能转化为人们可以直接阅读和理解的信息。这种针对事物整体数据的统计处理工作，被称为描述统计。

3. 推断统计

在有些情况下，人们获得的统计资料并非事物整体的状况，而是来自事物的一个局部。如果利用局部的数据去推断整体的情况，以及这种推断的有效性和可靠性如何，即是推断统计所要研究的内容。

4. 多元统计分析

在统计课程设计中，多元统计分析是一个独立的部分，主要涉及到对多变量情况的研究。例如，描述一个人的能力，需要从科研能力、动手能力、组织能力等多个方面进行综合判断，如果对涉及多个变量的统计问题进行研究，即为多元统计的内容。多元统计根据掌握信息的不同，也可分为多元描述统计和多元推断统计，但基本方法大多需要涉及到矩阵等工具，属于统计学中要求较高的部分。

1.2.3　统计数据

1. 数据的计量尺度

统计数据是对客观现象进行计量的结果。对客观现象进行计量，就必须弄清楚数的计量尺度问题。根据对研究对象计量的不同精确程度，将计量尺度由低到高、由粗略到精确分为四个层次：定类尺度、定序尺度、定距尺度和定比尺度。

(1)定类尺度。定类尺度亦称为列名尺度，是最粗略、计量层次最低的计量尺度。它是按照客观现象的某种属性对其进行平行的分类。此时，若用数字表示，该数字仅作为各类的代码，度量各类之间的类别差，不反映各类的优劣、量的大小或顺序。例如，人口按性别分为男女，用“1”表示男性，用“0”表示女性。定类尺度的主要数学特征是“＝”或“≠”。在统计处理中虽然可以计算单位数，但它不能表明第一类的一个单位可以相当于第二类的几个单位。

(2)定序尺度。定序尺度亦称为顺序尺度，它是对客观现象各类之间的等级差或顺序差的一种测度，是比定类尺度更高一级的计量尺度。定序尺度不仅可以将研究对象分成不同的类别，而且还可以反映各类的优劣、量的大小或顺序。例如，学生成绩可以分为优、良、中、及格和

不及格等五类，在这里，定序尺度虽然无法表明一个优等于几个良，但却能确切地表明优高于良，良又高于中。定序尺度的主要数学特征是“＜”或“＞”。在统计的变量数列中可以确定其中位数、分位数等指标的位置。

(3)定距尺度。定距尺度亦称为间隔尺度，它是对现象类别或次序之间间距的测度，是比定序尺度更高一级的计量尺度。定距尺度不但可以用数表示现象各类别的不同和顺序大小的差异，而且可以用确切的数值反映现象之间在量方面的差异。定距尺度使用的计量单位一般为实物单位(自然或物理)或者价值单位。反映现象规模水平的数据必须以定距尺度计量，例如，产品产量、人口数、企业数、国内生产总值、气象的温度湿度及各种试验数据都以定距尺度为计量尺度。定距尺度的主要数学特征是“＋”或“－”。定距尺度在统计数据中，占据重要的地位，统计中的总量指标就是运用定距尺度为计量尺度。

(4)定比尺度。定比尺度亦称为比率尺度，它是在定距尺度的基础上，确定相应的比较基数，将两种相关的数加以对比而形成相对数(或平均数)，用于反映现象的结构、比重、速度、密度等数量关系。例如，将一国的国内生产总值与该国的人口数对比，计算人均国内生产总值，以此反映该国的经济能力。定比尺度的主要数学特征是“×”或“÷”。在统计的对比分析中，广泛地运用定比尺度进行计量。

2. *数据的类型*

数据是对现象进行测量的结果。例如，对经济活动总量的测量可以得到国内生产总值(GDP)数据，对股票价格变动水平的测量可以得到股票价格指数的数据，对人口性别的测量可以得到男或女这样的数据，等等。由于使用的测量尺度不同，统计数据可以分为不同的类型。下面从不同角度说明数据的分类。

(1)按照所采用的计量尺度不同，可以将数据分为分类数据、顺序数据和数值型数据。

定义 1.2 分类数据：只能归于某一类别的非数字型数据。

分类数据是对事物进行分类的结果，数据表现为类别，是用文字来表述的。它是由定类尺度计量形成的。例如，人口按性别分为男、女两类，企业按照经济性质分为国有、集体、私营、合资、独资企业等，这些均属于分类数据。为便于统计处理，对于分类数据，可以用数字代码来表示各个类别，例如，用 1 表示“男性”，0 表示“女性”等。

定义 1.3 顺序数据：只能归于某一有序类别的非数字型数据。

顺序数据也是对事物进行分类的结果，但这些类别是有顺序的。它是由定序尺度计量形成的。例如，将产品分为一等品、二等品、三等品、次品等；考试成绩分为优、良、中、及格、不及格等；一个人对某一事物的态度分为非常满意、满意、保持中立、不满意、非常不满意等。同样，对顺序数据也可以用数字代码来表示，例如，1 表示非常满意，2 表示满意，3 表示保持中立，4 表示不满意，5 表示非常不满意。

定义 1.4 数值型数据：按数字尺度测量的观察值。

数值型数据是使用自然或度量衡单位对事物进行测量的结果，其结果表现为具体数值。现实中处理的大多数都是数值型数据。

(2)按照数据的收集方法，可以将其分为观测数据和实验数据。

定义 1.5 观测数据：通过调查或观测而收集到的数据。

观测数据是在没有对事物人为控制的条件下而得到的，有关社会经济现象的统计数据几乎都是观测数据。

定义 1.6　实验数据：在实验中控制实验对象而收集到的数据。

例如，对一种新药疗效的实验数据，对一种新的农作物品种的实验数据。自然科学领域的大多数数据都为实验数据。

(3)按照被描述的对象与时间的关系，可以将统计数据分为截面数据和时间序列数据。

定义 1.7　截面数据：在相同或近似相同的时间点上收集的数据。

截面数据所描述的是现象在某一时刻的变化情况。例如，2012 年我国各地区的国内生产总值数据就是截面数据。

定义 1.8　时间序列数据：在不同时间上收集到的数据，称为时间序列数据。

时间序列数据所描述的是现象随时间而变化的情况，例如 2003 年至 2012 年我国的国内生产总值数据就是时间序列数据。关于时间序列，本书将在第 9 章中进行详细介绍。

3. *数据的表现形式*

统计数据通常表现为：绝对数、相对数和平均数。

(1)绝对数。现象的规模、水平一般以绝对数形式表现，例如，国内生产总值、人口数、进出口额等。绝对数的计量单位一般为实物单位或价值单位，有时也采用复合单位。实物单位可以是自然计量单位，也可以是物理计量单位，如人口数用人计量，粮食产量用吨计量，耕地面积用公顷计量等，对于一些化工产品，常常折合成为标准实物单位。价值单位是以货币形式进行计量，如国内生产总值、进出口总额是以价值单位为计量单位。复合计量单位是由两种或两种以上计量单位复合而成的，如以“吨公里”为货物周转量的计量单位，以“千瓦时”为用电量的计量单位。

绝对数按其反映的时间状态不同，分为时期数据和时点数据。时期数据是反映现象在一段期间内发展过程的总量，它具有连续统计和可加性的特点，其数值大小与所属的时间长短有直接关系，如国内生产总值、进出口总额。时点数据是反映现象在某一特定时点所处的状态，它是采用间断登记方式取得资料的，不具有可加性，其数值大小与时点间隔长短没直接关系，如期末人口数、期末在建工程投资额等。

(2)相对数。相对数是由两个绝对数对比而得的，常用的相对数有：结构相对数、动态相对数、比较相对数、比例相对数、强度相对数、利用程度相对数、计划完成相对数等。相对数的计量单位大部分是无名数，但也有一些是采用有名数为计量单位。把对比基数抽象为 100 而计算的相对数为百分数，把对比基数抽象为 1000 而计算的相对数为千分数，这些都是无名数。如果把某地区的人口数与该地区的土地面积对比所计算的相对数是一种强度相对数，被称为人口密度相对数，其计量单位为“人/平方公里”，这就是有名数。

(3)平均数。统计平均数是用于反映现象总体的一般水平或分布的集中趋势，数值平均数是由总体标志总量对比总体单位数而计算的。关于这部分内容，本书将在第 3 章中进行介绍。

1.3　统计学的基本概念

1.3.1　总体和样本

1. *总体*

定义 1.9　总体：总体是由客观存在的、具有某种共同性质又有差别的许多个别单位所构

成的整体，当这个整体作为统计研究对象时称统计总体，简称总体。

例如，研究某个工业部门的企业生产情况时，该部门的所有工业企业可以作为一个总体，因为它是由许多客观存在的工业企业组成的，而每个工业企业都是进行工业生产活动的基层单位，具有同质性。

如果一个统计总体中包括的单位数是无限的，称为无限总体，例如，连续大量生产某种零件时，其总产量是无限的，构成一个无限总体。总体中包括的单位数是有限的，称为有限总体。例如，在特定时点上的人口总数、工业企业总数等，都是有限总体。对于有限总体，既可以进行全面调查，也可以抽样调查。对于无限总体来说，只能进行抽样调查，根据样本数据推断总体特征。此外，统计总体还可以分为静态总体和动态总体，前者所包含的各个单位属于同一个时间，后者所包含的各个单位则属于不同时间。根据一定的目的，针对这两类总体就可以分别进行静态研究或动态分析。

综上所述，可见总体和总体范围的确定、取决于统计研究的目的要求。而形成统计总体的必要条件，亦即总体必须具备三个特性：大量性、同质性和差异性。

(1)大量性。大量性是总体的量的规定性，即指总体的形成要有一个相对规模的量，仅仅由个别单位或极少量的单位不足以构成总体。因为个别单位的数量表现可能是各种各样的，只对少数单位进行观察，其结果难以反映现象总体的一般特征。统计研究的大量观察法表明，只有观察足够多的量，在对大量现象的综合汇总过程中，才能消除偶然因素，使大量社会经济现象的总体呈现出相对稳定的规律和特征，这就要求统计总体必须包含足够多数的单位。足够多数，是指足以反映规律的数量要求。当然，大量性也是一个相对的概念，它与统计研究目的、客观现象的现存规模以及总体各单位之间的差异程度等都有关系。

(2)同质性。总体的同质性，是指构成总体的各个单位至少有一种性质是共同的，同质性是将总体各单位结合起来构成总体的基础，也是总体的质的规定性。例如，全国工业企业作为统计总体，则每个总体单位都必须具有从事工业生产活动的企业特征，而不具有这些特征的就不能称之为工业企业。如果违反同质性，把不同性质的单位结合在一起，对这样的总体进行统计研究，不仅没有实际意义，甚至会产生虚假和歪曲的分析结论。

同质性的概念是相对的，它是根据一定的研究目的而确定的，目的不同，同质性的意义也就不同。例如，研究全国工业企业的生产状况时，所有工业企业都是同质的，而研究民营工业企业生产状况时，那么，民营工业企业与国有工业企业就是异质的。可见，同质性是相对研究目的而言的，当研究目的确定后，同质性的界限也就确定了。

(3)差异性。总体各个单位除了具有某种或某些共同的性质以外，在其他方面则各不相同，具有质的差别和量的差别，这种差别称为变异。正因为变异是普遍存在的，才有必要进行统计研究，是统计的前提条件。总体中各个单位之间具有变异性的特点，这是由于各种因素错综复杂作用的结果，所以有必要采用统计方法加以研究，才能表明总体的数量特征。

定义 1.10 总体单位：构成总体的每一个事物或基本单位。

原始资料最初就是从各个总体单位取得的，所以总体单位是各项统计数字最原始的承担者。例如，研究某个工业部门的生产情况时，该工业部门的所有工业企业可以作为一个总体，每个工业企业则是总体单位，将每个工业企业的某些数量特征加以登记汇总，就取得该工业部门的统计资料。

总体和总体单位是相对而言的，在一次特定范围、目的的统计研究中，统计总体与总体单位是不容混淆的，二者的含义是确切的，是包含与被包含的关系。但是随着统计研究目的及范围

的变化，统计总体和总体单位可以相互转化。同一事物在不同情况下，可以作为总体，也可以作为总体单位。例如，在上述某一工业部门所有工业企业的统计总体中，每个企业是一个总体单位。但为了要研究一个典型企业的内部问题时，则被选作典型的某一企业又可作为一个总体。

2. 样本

定义 1.11　样本：从总体中抽取的一部分元素所组成的集合称为样本。构成样本的元素数目，称为样本容量。

例如，一家公司正在接受审计，审计人员没有必要对该公司年度内的所有 55400 张发票全部审查，只需随机抽查一个 100 张发票的样本即可，审计人员通过这 100 张发票计算的差错率可对全部 55400 张发票的差错率进行推断，其中 100 即为样本容量。

样本是从总体抽取出的、作为总体的代表、由部分单位组成的集合体。在抽样推断中，总体又被称为母体，相应地，样本也被称为子样。抽取样本应注意如下几个问题：

(1)样本的单位必须取自总体，这是因为抽取样本的目的是为了推断总体，所以，不允许以总体外部的单位作为该总体的样本；

(2)一个总体可以抽取许多样本，样本个数的多少与抽样方法有关；

(3)样本的抽取必须排除主观因素的影响，以确保样本的客观性与代表性。

1.3.2　标志和指标

1. 标志

定义 1.12　标志：用来说明总体单位属性和特征的名称。

标志按其表现形式有数量标志与品质标志两种。凡是表示总体单位数量特征的标志，称数量标志。它能用数量来表示，如企业的职工人数、产量、产值；职工的年龄、工龄、工资等。凡是表示总体单位“质”的特征的标志，称品质标志。如职工的性别、企业的经济类型、工人的工种等。标志的具体表现是在标志名称之后所表明的属性或数值，如某职工的性别是女，民族是汉族。这里的“性别”和“民族”是品质标志名称。而“女”和“汉族”是这类标志的属性的具体表现。又如该工人的年龄是 35 岁，工资是 1960 元，则“年龄”和“工资”是数量标志的名称，而“35 岁”和“1960 元”则是它们的数值表现。

按照标志在总体中各单位的具体表现是否相同，标志可分为不变标志和可变标志。标志在同一个总体中的各单位具体表现都相同，我们称之为不变标志。例如，在教师总体中，职业这一标志在各单位的表现都是相同的，都是教师，在此，职业就是不变标志。一个总体中，至少要有一个不变标志，才能把各单位结合成为一个总体。如果没有不变标志，那么，总体也就不存在。由此可见，不变标志是总体同质性的基础。标志在同一个总体中的各单位具体表现有可能不同，我们称之为可变标志。可变标志的属性或特征的具体表现是由一种状态变为另一种状态，统计上称之为变异，因而可变标志也称为变异标志。例如，在教师总体中，教龄这一标志在各单位的表现可能不同，在此，教龄就是可变标志。在一个总体中，如果不存在可变标志，或者说所研究的现象总体在各单位之间不存在着任何差异，这就无须做调查，无须进行统计研究了。所以，总体的同质性是问题研究的基础，而总体的差异性则是问题研究的本质。

2. 指标

定义 1.13　指标：反映统计总体的数量特征的概念和数值称为指标。

一个完整的统计指标是由两个部分所构成，即指标名称和指标数值。指标名称和指标数值是两个既有联系又有区别的概念。指标名称是统计所研究的社会经济现象的科学概念，表明社会经济现象的质的规定，反映某一社会现象内容所属的范围；指标数值则是统计所研究现象的具体数量综合的结果，对某一社会经济现象总体特征从数量上加以说明。统计指标名称及其指标数值的有机结合，也就是事物质的规定性和量的规定性有机联系的表现。

指标一般包含有六个要素：即指标名称、计量单位、核算方法、时间限制、空间限制和指标具体数值。例如，我国 2011 年国内生产总值为 473104 亿元。该统计指标就包含上述六个要素。

表 1—1　统计指标要素举例

指标名称	计量单位	核算方法	时间限制	空间限制	指标具体数值
国内生产总值	亿元	生产法、收入法和支出法	2011	中国	473 104

资料来源：国家统计局《中国国内生产总值年度核算说明》

统计指标按其所反映的数量特点和内容的不同，可以分为数量指标和质量指标两类。

凡是反映社会经济现象范围的广度、规模大小和数量多少的指标叫数量指标，它表示事物外延量大小。例如人口总数、企业总数、耕地面积、工业总产值和商品流转额等，都属于这一类指标。数量指标是用绝对数表示的，并具有实物的或货币的计量单位。统计实践中这类指标通常是以总量指标的形式出现。由于数量指标反映的是现象总体的绝对量，因而其指标数值大小随总体范围的大小而增减变动。

反映现象本身质量、现象的强度、经营管理工作质量和经济效果等的统计指标，称为质量指标，它表示事物的内涵量状况。例如产品合格率、固定资产的利用程度、单位成本指标、利润率、劳动生产率，等等。质量指标是用相对数或平均数表示的，统计工作中，这类指标通常是以相对指标或平均指标的形式出现。由子质量指标反映的是现象总体内部的数量关系，因而其指标数值大小与总体范围大小没有直接的关系。数量指标和质量指标的关系表现在，数量指标是计算质量指标的基础，质量指标往往是相应的数量指标进行对比的结果。

3. 标志与指标的关系

(1)指标与标志的区别：

第一，反映的对象和范围大小不同。统计指标说明的是总体的数量特征，而标志则是反映总体单位的数量特征。

第二，表述形式不同。统计指标都可以用数值表示，而标志既有能用数值表示的数量标志，又有不能用数值只能用文字表述的品质标志。

(2)指标与标志的联系：

第一，具有对应关系。在统计研究中，标志与统计指标名称往往是同一概念，具有相互对应关系。因此，标志就成为统计指标的核算基础。

第二，有汇总关系。有许多统计指标的数值是从总体单位的数量标志值汇总而来的，如某地区工业总产值就是各企业总产值加总之和，这里，地区工业总产值就是统计指标，而各企业总产值则是标志。同时，通过对品质标志的标志表现所对应的总体单位数进行加总，也能形成统计指标。例如上述的工业企业经济类型，汇总后可得出具有某种属性的总体单位数，如国有经济企业数、集体经济企业数等。

第三，指标与数量标志之间存在着变换关系。由于研究的目的不同，统计总体和总体单位具有相对性。随着研究目的的变化，原来的统计总体如果变成总体单位，则相对应的统计指标

也就变成数量标志，反之亦然。

1.3.3　参数和统计量

定义 1.14　参数：用来描述总体特征的概括性数字度量，是研究者想要了解的总体的某种特征值。

我们所关心的参数通常有总体平均数、标准差、总体比例等。由于总体数据通常是不知道的，所以参数是一个未知的常数。例如，我们不知道一个地区所有人口的平均年龄，不知道一个城市所有家庭的收入差异，不知道一批产品的合格率等。正因为如此，我们才进行抽样，根据样本计算出某些值去估计总体参数。

定义 1.15　统计量：用来描述样本特征的概括性数字度量，是根据样本数据计算出来的一个量，它是样本的函数。

通常我们所关心的统计量有样本平均数、样本方差、样本比例等。由于样本是我们已经抽取出来的，所以统计量总是已知的。抽样的目的就是要根据样本的统计量去推断总体的参数。例如，用样本平均数去估计总体平均数，用样本标准差去估计总体标准差，用样本比例去估计总体比例，等等。

1.4　统计学的应用

1.4.1　统计学的应用领域

说出哪些领域应用统计，这很困难，因为几乎所有的领域都应用统计；要说出哪些领域不使用统计，同样也很困难，因为几乎找不到一个不用统计的领域。可以说，统计是适用于所有学科领域的通用数据分析方法，是一种通用的数据分析语言。只要有数据的地方就会用到统计方法。这里，我们不想列举统计的应用领域，只想通过几个简单的例子说明统计的应用。

案例 1：为了调查研究“人在戒烟后体重会增加”这一断言，研究人员选择了一个由 400 个参与者构成的样本，他们都成功地参与了戒烟运动。每个人在活动开始前和一年后都称量了体重。参与者体重的平均变化是增加了 5 磅。研究人员由此总结说有证据表明这一断言是正确的。

案例 2：在统计学应用的诸多领域中，文学著作的统计分析是一个饶有兴趣的分支。《红楼梦》是我国四大名著之首，而且有很多悬而未决的问题，把统计学的定量分析方法引入红学研究是很自然的。早在 1980 年，在美国威斯康星大学召开的“首届国际《红楼梦》研讨会”上，该校华裔学者陈炳藻教授首次报告了他在这方面的研究工作，此后还出版了专著。陈教授将《红楼梦》120 回分为三组，每组 40 回，并将《儿女英雄传》作为对照组进行比较研究。他从每组中任取 8 万字，挑出名词、动词、形容词、副词、虚词这 5 种词，然后运用统计学方法算出各组之间用词的相关程度，结果发现：《红楼梦》前 80 回与后 40 回所用词汇的相关程度远远超过《红楼梦》与《儿女英雄传》所用词汇的相关程度，并由此推断：前 80 回与后 40 回均为曹雪芹一人所作。我国华东师范大学陈大康教授得出了迥异的结论，他也把《红楼梦》120 回分成三组，每组 40 回，并统计了其中所含词、字、句等 88 个项目。他发现，这些词在前两组出现的规律相同，而与后 40 回却不一致；由此推断：后 40 回非曹雪芹所作（但含有少量残稿）。复旦大学李贤平教授又提出“成书新说”。李教授选择了 47 个虚字为识别特征，诸如：“之、其、或、亦、了、的、不、把、别、好”

等，利用各种统计方法（主成分分析、典型相关分析、聚类分析等），对它们在书中各回的出现频率进行统计分析，探索各回写作风格的接近程度，并用三个层次的聚类方法对各回进行分类。由此提出了成书过程新观点：《红楼梦》前 80 回是曹雪芹根据《石头记》增删而成；而后 40 回则是曹家亲友搜集整理原稿加工补写而成。

案例 3："挑战者号"航天飞机失事预测。1986 年 1 月 28 日清晨，载有 7 名宇航员的"挑战者号"进入发射状态。就在发射前，有冰片牢附在机壳上。几分钟后，正当电视新闻报道它已进入轨道时，航天飞机在毁灭性的爆炸声中化成碎片，机上的宇航员片骨未存。推动航天飞机进入太空的两个固体燃料发动机是由 Thiokol 公司制造的。失事前一天晚上，Thiokol 公司的经理们和美国宇航局（NASA）就如期发射还是推迟发射产生了争执。天气预报发射时的气温为 31°F。争执的结果是采纳了 Thiokol 公司经理们的建议：按计划发射航天飞机，因为他们觉得没有确凿证据表明低温会对固体燃料火箭推进器的性能产生影响。在此次失事前，该航天飞机 24 次发射成功。将航天飞机送入太空的两个固体燃料推进器有 6 只 O 型项圈密封。在几次飞行中，曾发生过 O 型项圈被腐蚀或气体泄漏事故。这样的事故是极其危险的，前 24 次发射中有一次发动机遭到了永久性破坏。根据 23 次飞行中发生腐蚀或泄漏事故的次数（因变量 y）及火箭连接处的温度（自变量 x）数据，进行线性回归得到的回归方程为 $\hat{y} = 3.698 - 0.04754x$。当温度为 31°F，O 型项圈发生事故的预计次数为 2.225 次。结果显示连接处的温度与 O 型项圈事故之间有一定的相关性。如果当时那些经理们看到了回归的预测结果，也许推迟发射会成为其谨慎的选择。

前两个是统计得以应用并取得成效的例子，后一个是统计结果未被采纳而酿成惨剧的例子。不管怎样。它们都表明统计在许多领域都有广泛应用。

1.4.2 统计的误用与滥用

大约一个世纪以前，政治家 Benjamin Disraeli 曾有一个著名的论断："有三类谎言：谎言、糟透的谎言和统计。"他还说："图没有说谎，是说谎者在画图。"历史学家 Andrew Lang 说："一些人使用统计就像喝醉酒的人使用街灯柱——支撑的功能多于照明。"这些叙述是指，在统计的使用中，数据以设计好的误导方式被提供。一些统计的滥用者只是简单的无知或粗心大意，但其他人却怀有某种个人目的，希望隐藏以对己不利的数据而强调有利的数据。我们现在就将提供一些例子，其中数据的提供方式可能被歪曲了。统计学的滥用主要来源以下几个方面：

1. 不好的样本

这是滥用统计的一个主要来源：就是指使用不恰当的方法收集数据。瑞士心理学家 H. C. Lombard 有一次通过参考包括姓名、死亡时的年龄以及职业的死亡证明，汇集了不同职业的寿命数据，然后他计算出不同职业的平均寿命，发现学生排在最后，平均只有 20.7 岁，这里，问题就出在数据是不恰当的，因为大多数人在相对年轻时都会是学生；而当他们年龄渐长时就不再是学生了，而是从事了其他职业。因此，20.7 岁这个寿命数字只能说明，当学生死亡时，他们可能很年轻。如果你计算十几岁的死亡者的平均年龄，也会犯同样的错误，该数字不可能超过 19 岁，而从事试飞员职业的人死亡时的平均年龄肯定大大高于 19 岁，但这并不能真的意味着做一个十几岁的人就比做一名试飞员更危险。

2. 小样本

基于非常小的样本而得出广泛的结论或推论是很容易让人误解的。例如，由维护儿童权益

基金会出版的《美国的校外儿童》中说，在一个地方被暂停学籍的中学生中，67％的学生至少被暂停了 3 次。这个数字是在一个仅有 3 个学生的样本基础上得出的，而媒体却没有说明样本的规模这么小。有的时候一个样本看起来相对比较大（例如一个对“2000 个随机选择的成年美国人”进行的调查），但如果结论只是在其中的一小群的基础上得出的，例如缅因州的男性天主教徒的共和党人，这样的结论也可能是在太小的样本上得出的，尽管有一个足够大的样本很重要，但大小并不是唯一的问题；样本数据以一种恰当的方式收集同样重要，例如通过随机选择。有时，即使是的大的样本也可能是不好的样本。

3. 带有引导性的问题

调查中问题的措辞可能会引出想得到的回答。在最近一次调查中，问题的不同措辞就导致了不同的回答：例如一项来自德国的民意测验中的这些出现这样的问题：

你认为汽车对空气污染的作用比工厂大还是小？（63％谴责汽车）；

你认为工厂对空气污染的作用比汽车大还是小？（61％谴责工厂）。

4. 误导性的图表

不当的使用图表可能会夸大或减小数据的真正含义。图 1－1 描绘了 2008 年全职男女性工资的周平均收入，但是图(a)的设计就夸大了男性和女性工资之间的不同。由于图(a)的水平轴并没有从零开始，它就给人一种错误的主观印象。图 1－1 带给我们一个重要的教训：想要正确理解一个图，我们必须分析这幅图所提供的数量方面的信息，这样才不会被图的大体形状所误导。

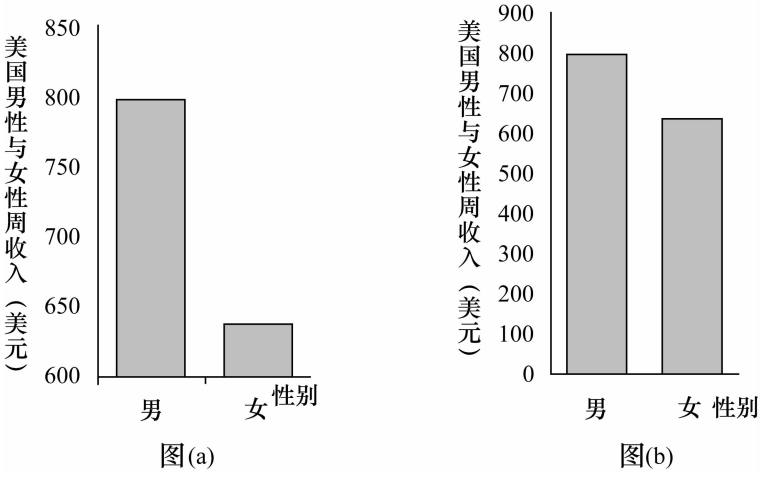

图 1－1　2008 年美国全职男女性周平均收入图

5. 局部描述

NBA 深锐观察是新浪 NBA 的一个著名专栏，在这一期“深锐观察”：无三分不冠军？湖人痼疾不除只能等死”这一期中，作者列举了“NBA 历年季后赛三分球数据”，其中包括“总冠军出手次数”、“联盟平均出手次数”、“总冠军命中率”、“联盟平均命中率”四项。作者在文中声称三分球的命中率是最后的总冠军球队一个重要的指标，但是作者仅仅比较了表格的前几行与最后几行，仔细观察不难发现，2010、2006、1998、1997、1996 等多个年份的数据都不能支持作者的结论，因而他运用了局部描述的方法，让不细看表格的读者相信作者的观点是有数据支撑的。

6. 百分数的陷阱

美国多克斯牙膏公司做过一个用户调查，调查结果显示使用多克斯(Doakes)牌牙膏将使蛀牙减少 23%！大字标题历历在目。这些结论出自一家信誉良好的“独立”实验室，并且还经过了注册会计师的证明。似乎是值得信任的。但是事实上读者只有在读小字的内容时才会发现，实验的样本仅由 12 人组成。这全然不能说明多克斯牙膏的神奇功效。

此外，统计也往往被作为两个极端使用：一个极端是不懂或不太懂统计的人认为统计没什么用。他们因为不懂统计而瞧不起统计，他们不用或几乎不用统计方法分析数据，即使作些统计分析，也往往是表面上的、走入这一极端的人，他们的决策依据是自己的大脑。另一个极端是把简单问题复杂化，特别是在管理领域，一些管理者把本来可以简单方法解决的问题故意复杂化，他们不用简单的分析方法，而是用复杂的分析方法；他们为证明管理的科学性，建立一个别人看不懂的模型，编一大堆程序，输出了一大堆数字和符号；他们得出用统计语言陈述的结论，提出一些似是而非的建议；等等。这样的分析往往是脱离了管理问题，对实际决策也未必有用。在统计应用中，这两个极端都是不可取的。管理决策中不用统计几乎不可想象；把简单问题复杂化对管理决策也未必有用。

统计是一门科学，也是一个工具，可以用来反映经济和社会发展状况，但是，目前社会上出现了一种以误读和曲解统计数据以吸引公共眼球的现象，这是违背科学的和不负责任的，统计应该恰当地应用到它能起作用的地方。不能把统计神秘化，更不能歪曲统计，把统计作为掩盖事实的陷阱。

第 2 章　统计资料的搜集与整理

2.1　统计资料的搜集

人们要从数量上认识客观对象，就必须通过调查或实验来搜集数据。要搜集什么数据，采用哪种方法去搜集数据，才能保证数据的搜集工作适时、有效地进行，这就需要研究搜集资料的理论与相应的方法。

统计数据搜集的方式有两种：一种是直接向调查对象搜集反映调查单位的统计资料，一般称为原始资料，又称为初始资料；另一种是根据研究的目的，搜集已经加工整理过的、说明总体现象的资料，一般称为次级资料或第二手资料。本节研究的是原始资料搜集的理论与方法，对原始资料的搜集亦称为统计调查。

2.1.1　统计调查的种类

1. 普查

普查是专门组织的一种全面调查，它主要是用以调查某些不能或不宜用定期的全面报表搜集的统计资料。借助普查可系统地、全面地掌握一个国家（或地区）的人力资源、财力资源和物资资源的数量、分布及利用状况。

普查有两个主要特点，第一，它是非经常性的调查，一般间隔较长的时间才进行一次。第二，它是一种全面调查，它比任何一种调查形式更能掌握大量、详细、全面的统计资料。

普查的组织形式有两种，一种是通过组织普查机构，配备一定数量的普查人员，对调查单位直接进行登记调查，如我国人口普查就是采用这种形式。另一种是利用调查单位的原始记录和核算资料，结合清库盘点，由调查单位自行填报调查表格，如我国物资库存普查就是采用这种形式。

2. 抽样调查

抽样调查是按随机原则从调查对象中抽取一部分单位作为样本进行观察，然后根据所获得的样本数据，对调查对象总体特征做出具有一定可靠程度的推算。

抽样调查是科学研究和现代管理决策的一种重要调查方法，是一种既节省人、财、物力又能保证一定可靠性的科学方法。国家统计系统设有城调队和农调队，就是利用抽样调查方法搜集城市居民收支和消费、农业产量和农民生活的有关数据。抽样调查方法在企业产品质量控制和检验等方面均有广泛的应用，该方法的介绍请阅第 5 章抽样与抽样分布。

3. 重点调查

重点调查是指在调查对象中，只选择一部分重点单位进行的非全面调查。所谓重点单位，是针对现象量的方面而言，尽管这些单位在全部单位中只是一部分，但它们在所研究现象的标志总量中却占有很大的比重，在总体中具有举足轻重的作用，因而对这些单位的调查就能够从数量上反映整个总体在该标志总量方面的基本情况。重点单位的确定，是组织重点调查的一个重要问题，重点单位的选择始终着眼于它在所研究现象的标志总量中所占的比重。重点单位可能是一些企业，也可能是一些地区、城市。

重点调查由于选择的单位较少，因而调查项目就允许多一些，所了解的情况也可以详细一些。一般地说，当调查任务只要求掌握基本情况，而部分单位又能比较集中地反映研究的项目时，采用重点调查比较适宜。

4. 典型调查

典型调查是一种专门组织的非全面调查。它是根据调查的目的，在对所研究的对象进行初步分析的基础上，有意识地选取若干具有代表性的单位进行调查和研究，借以认识事物发展变化的规律。

典型调查具有灵活机动，通过少数典型即可取得深入、翔实的统计资料的优点。但是，这种调查由于受“有意识地选出若干有代表性”的限制，在很大程度上受人们主观认识的影响。因此，必须同其他调查方法结合起来使用，才能避免出现片面性。

5. 统计报表制度

统计报表是按照国家统一规定的表格形式、统一规定的指标内容、统一规定的报送程序和报送时间，由填报单位自下而上逐级提供统计资料的一种统计调查方式。

国家利用统计报表定期地取得全社会的国民经济与社会发展情况的基本统计资料，是国家取得调查资料的主要方法之一。它已形成一种制度即统计报表制度。执行统计报表制度，是各地区、各部门、各基层单位必须向国家履行的一种义务。

统计报表制度的内容包括以下几个方面：

(1)表式。它是由国家统计部门根据研究的任务与目的而专门设计制定的统计报表表格，用于搜集统计资料。它是统计报表制度的主体。

(2)填表说明。它是对统计报表的统计范围、指标等作出的规定。

(3)指标解释。对列入表的统计指标的口径、计算方法以及其他有关问题的具体说明。

(4)分类目标。有关统计报表主栏中应进行填报的有关项目的分类。

(5)其他有关事项的规定。除了以上各项规定以外的一些注意事项，如：报送日期，报送方式，报送分数等。统计报表的资料来源，主要是基层的原始记录、台账及基层的内部报表。

2.1.2 统计调查的方法

1. 观察法

观察法是指调查人员不直接与受访者进行接触，而是通过旁观的方法获得对受访者情况的了解。观察法中，调查人员亲自到现场对调查对象进行观察计量取得资料，一般资料准确，但人

力多、时间长。

观察法一般用于对受访者客观状况进行调查，例如通过观察普通消费者在超市中选购商品的过程，可以分析出消费者对商品各方面属性的偏好情况。在使用观察法时，要求访问员具有较强的观察能力和心理分析能力，能够敏锐地发现受访者的各种无意识活动。

2. 采访法

采访法是指调查人员根据访问提纲，与受访者进行交谈，由此获得对受访者情况的了解。

在使用采访法时，访问员需要及时掌握受访者的谈话内容，对于有价值的信息进行深入追问。

采访法能够发现受访者的许多深层次的主观意见，因而常用于深度分析。但采访法的效果受访问员个人能力的影响很大，而且受访者的谈话漫无边际，很难进行定量分析。

3. 报告法

报告法是指由受访者填写有关报告表格，向调查人员报告自身情况。报告法是我国政府统计的传统方法，尤其是在计划经济时代，政府统计信息主要来自于各行各业提供的统计报表。在组织良好的情况下，报告法能够在较低的成本下，快速地获得有关统计结果。但报告法受被调查机构的主观配合情况影响较大，在政府逐渐减少对企业的直接干预的情况下，报告法的应用受到了很大的限制。

4. 问卷调查法

问卷调查法是指调查人员利用格式化的调查问卷，向受访者进行询问。问卷调查法是目前最常用的调查方法，其优点在于利用问卷限定了访问员的询问方式和受访者的回答方式，从而有助于获得符合分析要求的定量数据。问卷调查法不需要访问员进行自由联想和发挥，从而降低了对访问员自身素质的要求，更适宜于大规模的民意和商业调查活动。

常见的问卷调查方法包括：

(1)入户访问；

(2)街头拦截式访问；

(3)电话调查；

(4)邮寄问卷调查；

(5)留置问卷调查；

(6)媒体问卷调查。

5. 网络调查法

网络调查是传统调查在新的信息传播媒体上的应用。它是指在互联网上针对特定的问题进行的调查设计、收集资料和分析等活动。与传统调查方法相类似，网络调查也有对原始资料的调查和对二手资料的调查两种方式。

互联网作为一种信息沟通渠道，它的特点在于开放性、自由性、平等性、广泛性和直接性等。由于这些特点，网络调查具有传统调查所不可比拟的优势：

(1)网络调查成本低。网络调查与面访、邮寄访问、电话访问等离线调查的根本区别在于采样方式不同。所以，网络调查的成本低主要指的是采样成本低。传统调查往往要耗费大量的人

力物力，而网络调查只需有一台上网的计算机，通过站点发布电子问卷或组织网上座谈，利用计算机及统计分析软件进行整理分析，省却了传统调查中的印刷问卷、派遣人员、邮寄、电话、繁重的信息采集与录入等工作与费用，既方便又便宜。据业内权威人士讲，根据经验，离线调查每一个样本的投人大概是 120 元—150 元，所以离线调查者在抽样时，总希望尽可能地减少样本数。当然其前提是所抽取的样本数必须能把调查误差控制在允许范围之内，从而有效地降低采样的成本。网络调查就没有这种顾虑。

(2)网络调查速度快。网上信息传播速度非常快，如用 E-mail，几分钟就可把问卷发送到各地，问卷的回收也相当快。利用统计分析软件，可对调查的结果进行即时统计，整个过程非常迅速，而传统的调查要经过很长一段时间才能得出结论。离线调查三种采样方式的速度存在差别，其中以面访最快，电话访问次之，邮件访问最慢。而网络调查采样所需的时间则要少得多，约需几天或十几天。如 2000 年 11 月 3 日，科技日报《中国将参与全球规模最大网络调查》一文称，“PlanetProject 网上民意调查活动将于 11 月 15～18 日举行，持续四天，通过八种语言进行。”中国互联网发展状况统计报告(1999 年 7 月)称，“CNNIC 在 1999 年 6 月 15 日至 1999 年 6 月 30 日期间进行了网上联机问卷调查”。由这些具体调查所花费的时间可以看出：网络调查采样速度远快于面访等离线调查。这与网络调查问卷的发放、填答、提交皆不受时空限制、即时快捷相关。无论是把问卷直接放在网上，还是发送 E-mail 或网上拦截，都可以迅速把问卷大范围地呈现在被访者面前。问卷的填答虽可能会费些时间，但填答时间由自己支配。填答完毕后，问卷的提交也比较简单，只要点击一下提交键即可。

(3)调查隐匿性好。在调查一些涉及个人隐私的敏感问题时，离线调查尽管可以在问卷设计中通过采用委婉法、间接法、消虑法、虚拟法等手段，在问题和被访者之间增加一些缓冲因素，但无论如何，离线调查各种采样方式都会在不同程度上影响到被访者的填答心理。一般而言，面访最大，电话访问次之，邮寄访问最小。而网民是在完全自愿的情况下参与调查，对调查的内容往往有一定的兴趣，因而回答问题时更加大胆、坦诚，调查结果可能比传统调查更为客观和真实。在网上“没人知道你是一条狗”，应该说，网络调查的隐匿性较离线调查高。网络调查的这一特点可使被访者在填答问卷时的心理防御机制降至最低程度，从而保证填答内容的真实性。

(4)网络具有互动性。网络调查的这一优势同样是基于网络自身的技术特性。网络的互动性赋予网络调查互动性的优势。网络调查不受时空的限制，可以 24 小时向天南海北、世界各地进行调查，抽样框相当大，调查范围也相当广泛。2000 年 11 月，3com 中国公司举行 Planet-Project 网上民意调查活动。这次活动中，PlanetProject 在自己的网址上设下不同题材的各种问题，围绕人类基本状况的方方面面展开。“让中国民众首次与来自世界其他地区、不同年龄和性别的人同时分享和比较彼此的观点和看法”。

根据调查方法的不同，网络调查可分为：E-mail 法、Web 站点法、Net-meeting 法、Internet Phone 等。

(1)E-mail 法。E-mail 法即电子邮件法，以较为完整的 E-mail 地址清单作为样本框，使用随机的方法发送问卷进行调查。

(2)Web 站点法。Web 站点法又称主动浏览访问法，即将调查问卷放置在访问率较高的 Web 站点的页面上，由对该问题感兴趣的访问者完成并提交。

(3)Net-meeting 法。Net-meeting 法即网络会议法,视频会议法,焦点团体座谈法。是通过直接在上网人士中征集与会者,并在约定时间举行网上座谈会,在主持人的引导下,对某一问题进行深入的或探索性的讨论、研究的一种网上调查方法。

(4)Internet Phone 法即网络电话法,是以 IP 地址为抽样框,采用 IP 自动拨叫技术,邀请用户参与调查。比如:可将 IP 地址排序,每隔 100 个进行一次抽样,被抽中的用户会自动弹出一个小窗口,询问其是否愿意接受调查,回答“是”,则弹出调查问卷;回答“否”,则呼叫下一个 IP 地址。这种调查方法类似传统调查方式中的电话调查。

网络调查对于调查工作者来说可能是一把双刃剑。一方面,网络调查的优势在于它可以在更广的范围内,对更多受众进行信息收集的工作。与传统研究方法相比,不仅是研究者可以以惊人的低价获得超乎想象之多的被调查者的情况和资料,网络调查还可以以自填的回答方式,通过标准化的方法向被调查者呈现一份多媒体的问卷,这显然是传统的调查方法难以做到的。但另一方面,网络调查潜在的危险是,日益增多的调查越来越良莠不齐,人们也难以区分好的调查与不好的调查。网络调查的价值也受到人们填答意愿的限制。因为在类似调查的狂轰乱炸下,人们可能干脆不理睬,也可能根据其内容、主题、娱乐性或者调查的其他特性而做出参与调查的决定,从而影响到网络调查的可信度。

2.1.3　统计调查方案的设计

经济管理人员和研究人员在实施调查之前,必须全面地计划,严密地组织事先要制定统计调查方案。所谓统计调查方案指关于统计调查的完整的工作计划,调查方案设计的好坏直接影响到调查数据的质量。

1. 确定调查目的

确定调查目的是调查方案设计中首先要解决的一个问题,调查目的是调查所要达到的具体目标,它所回答的是“为什么调查”,要解决什么样的问题等。

2. 确定调查对象和调查单位

调查对象是根据调查目的确定的调查研究的总体或调查范围。调查单位是构成调查对象中的每一个单位,它是调查项目和调查内容的承担者或载体,也是我们数据、分析数据的基本单位。

调查对象和调查单位所解决的是“向谁调查”,由谁来提供数据。例如,调查的目的是为了获取国有企业的资产负债分布状况,那么,所有的国有企业就是调查对象,而具体的每一个国有企业就是调查单位。在实际调查中,调查的单位可以是调查对象的全部单位,也可以是部分单位。如果采取全面调查方式,如普查,调查对象中的每一个单位都是调查单位;若采用非全面调查,如抽样调查,调查单位只是调查对象中的一部分单位。

3. 明确调查项目和制定调查表

调查项目要解决的问题是“调查什么”,也就是调查的具体内容。在拟定调查项目时应遵循以下原则:

(1)所列调查项目必须与统计研究目的一致,并且要“少而精”。

(2)所列调查项目必须通俗易懂,调查项目的提法要能使答案具有确定的表达形式,并且能够得到切实的答案。

(3)所列调查项目之间应该是一个有机的整体,以便使所收集的统计资料满足统计分析的需要,同时,也便于统计资料的检验,提高统计数据的质量。

在大多数统计调查中,调查项目通常以表格的形式来表现,称为调查表,它是用于登记调查数据的一种表格,一般由表头、表体和表外附加三部分组成。表头是调查表的名称,用来说明调查的内容、被调查单位的名称、性质、隶属关系等;表体是调查表的主要部分,它是调查内容的具体体现;表外附加通常由填表人签名、填表日期和填表说明等内容组成。

4. 确定调查时间

调查时间包含两方面的含义:一是调查资料所属的时间,它可以是一个时期,也可以是一个时点,它是由调查对象的特点决定的,二是调查工作的起止时间,它对保证调查工作的按期完成,是必要的。

5. 确定调查的组织实施计划

其内容包括确定调查机构,组织和培训调查人员,明确调查的方式、方法和进行调查的地点,落实调查经费的来源和开支办法,确定调查资料的报送方法和公布调查结果的时间等。

2.2 统计资料的整理与显示

数据搜集上来后,接下来就是数据的整理与显示的过程,使之符合统计分析的需要,同时对数据进行图表展示,以探索数据内在的数量规律性。统计数据的整理与显示,是实现统计研究的一个重要环节。统计整理与显示是统计调查的继续,也是统计分析的前提和基础条件,它具有承上启下的作用,成为人们对社会经济现象从感性认识上升到理性认识的过渡阶段。同时,统计整理质量如何,会直接影响统计分析的效果。

2.2.1 分类数据的整理与显示

分类的数据本身是对事物的一种分类,因而在整理时除了列出所分的类别外,还要计算出每一类别的频数、频率或比例、比率,同时选择适当的图形进行显示,以便对数据及其特征有一个初步的了解。

1. 频数与频数分布

定义 2.1 频数:也称次数,它是落在各类别中的数据个数。

我们把各个类别及其相应的频数全部列出,并用表格形式表现出来,就是频数分布。将频数分布用表格的形式表现出来就是频数分布表。

定义 2.2 比例:是一个总体中各个部分的数值占全部数值的比重,通常用于反映总体的构成或结构。

假定总体数量 N 被分成 K 个部分,每部分的数量分别为 $N_1, N_2, \cdots, N_K$,则比例定义为

$\frac{N_i}{N}$。各部分比例之和等于 1。

定义 2.3　百分比：将比例乘以 100 就是百分比或百分数。

例 2.1　一家市场调查公司为研究不同种类的饮料的市场占有率，对随机抽取的一家超市进行了调查。调查员在某天对 50 名顾客购买饮料的种类进行了记录，如果一个顾客购买某一种类的饮料，就将这一饮料的种类名称记录一次。求各种饮料购买数据的频数、比例、百分比。

表 2－1　顾客购买饮料种类记录表

茶类饮料	碳酸饮料	茶类饮料	果蔬汁饮料	乳饮料
乳饮料	茶类饮料	碳酸饮料	乳饮料	碳酸饮料
茶类饮料	碳酸饮料	碳酸饮料	功能饮料	茶类饮料
碳酸饮料	功能饮料	茶类饮料	碳酸饮料	功能饮料
功能饮料	乳饮料	乳饮料	功能饮料	乳饮料
碳酸饮料	茶类饮料	茶类饮料	果蔬汁饮料	果蔬汁饮料
果蔬汁饮料	茶类饮料	碳酸饮料	碳酸饮料	碳酸饮料
碳酸饮料	功能饮料	乳饮料	果蔬汁饮料	功能饮料
乳饮料	碳酸饮料	功能饮料	碳酸饮料	乳饮料
碳酸饮料	茶类饮料	功能饮料	果蔬汁饮料	茶类饮料

解：

表 2－2　饮料类型频数、比例、百分比表

饮料类型	频数	比例	百分比(%)
碳酸饮料	15	0.30	30
茶类饮料	11	0.22	22
功能饮料	9	0.18	18
果蔬汁饮料	6	0.12	12
乳饮料	9	0.18	18
合计	50	1	100

定义 2.4　比率：是各不同类别的数量的比值，可以是一个总体(或样本)中各不相同部分的数量对比。

如碳酸饮料与茶类饮料比率为 15∶11。为便于理解，通常将分母化为 1，如男女人数比率为 1.36∶1。比率由于不是总体(或样本)中部分与整体之间的对比关系，因而比值可能大于 1。

2. 分类数据的图示

(1)条形图。条形图是用宽度相同的条形的高度或长度来表示数据变动的图形，条形图可以横置或纵置，纵置时也称为柱形图，在表示分类数据的分布时，用条形图的高度(或长度)来表示各类别数据的频数或频率。

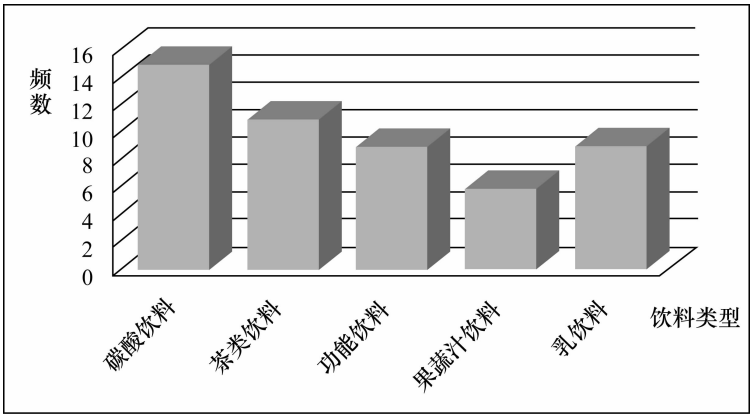

图 2—1　不同类型饮料的频数分布图

(2)圆形图(也称为饼图),是用圆形及圆内扇形的面积来表示数值大小的图形。主要用于表示总体中各组成部分所占的比例,对于研究结构性问题十分有用。总体中各部分所占百分比用圆内的各个扇形角度来表示。

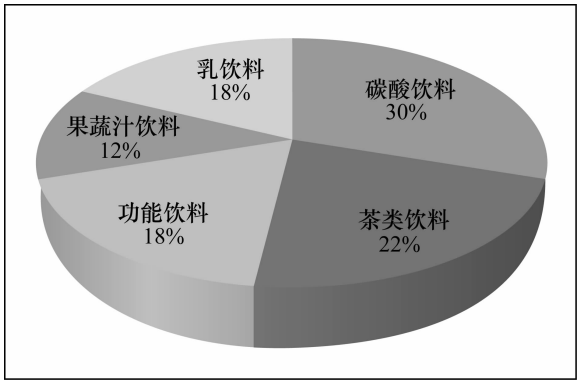

图 2—2　不同饮料类型构成图

2.2.2　顺序数据的整理与显示

分类数据的整理与显示方法,如频数、比例、百分比等都适用于对顺序数据的整理与显示。但有些方法适用于顺序数据,却不适用于分类数据。

1. 累计频数和累计频率

定义 2.5　累计频数:就是将各类别的频数逐级累加起来。一为向上累计:从类别顺序的开始一方向类别顺序的最后一方累加频数(数值型数据则从变量值小的一方向变量值大的一方累加频数);二为向下累计。

定义 2.6　累计频率:将各类别的百分比逐级累加起来,称为累计频率或百分比,也有向上累计和向下累计两种方法。

例 2.2　为评价家电行业售后服务的质量,随机抽取了由 100 个家庭构成的一个样本。服务质量的等级分别表示为:A. 好;B. 较好;C. 一般;D. 差;E. 较差。调查结果如下:

表 2—3　家电行业服务质量调查

B	E	C	C	A	D	C	B	A	E
D	A	C	B	C	D	E	C	E	E
A	D	B	C	C	A	E	D	C	B
B	A	C	D	E	A	B	D	D	C
C	B	C	E	D	B	C	C	B	C
D	A	C	B	C	D	E	C	E	B
B	E	C	C	A	D	C	B	A	E
B	A	C	D	E	A	B	D	D	C
A	D	B	C	C	A	E	D	C	B
C	B	C	E	D	B	C	C	B	C

要求制作一张频数分布表。

解:频数分布表如下:

表 2—4　甲地区消费者对家电行业售后质量评价的频数分布

服务质量等级	家庭数(户)	百分比(%)	向上累计		向下累计	
			户数(户)	百分比(%)	户数(户)	百分比(%)
A	14	14.00	14	14.00	100	100.00
B	21	21.00	35	35.00	86	86.00
C	32	32.00	67	67.00	65	65.00
D	18	18.00	85	85.00	33	33.00
E	15	15.00	100	100.00	15	15.00
合计	100	100.00	—	—	—	—

2. *顺序数据的图示*

(1)累计频数分布图。根据累计频数或累计频率,可以绘制累计频数分布或频率图。

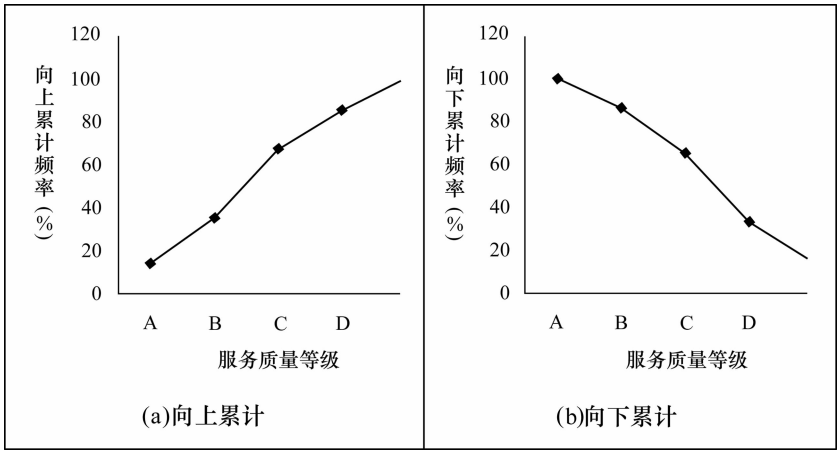

图 2—3　家电行业服务质量评价等级累计次数分布图

(2)环形图。环形图中间有一个“空洞”,总体中的每一部分数据用环中的一段表示。环形图与圆形图类似,但又有区别:圆形图只能显示一个总体各部分所占的比例;环形图则可以同时

绘制多个总体的数据系列，每一个总体的数据系列为一个环。环形图可用于结构比较研究，主要用于展示分类和顺序数据。

例 2.3 例 2.2 中为了得到更全面的数据，从另一地区“乙”随机抽取了由 100 家庭构成的一个样本二，调查他们对家电行业售后服务的质量的评价。得到的频数分布表如下：

表 2—5 乙地区消费者对家电行业售后质量评价的频数分布

服务质量等级	家庭数(户)	百分比(%)	向上累计		向下累计	
			户数(户)	百分比(%)	户数(户)	百分比(%)
A	16	16	16	16.00	100	100.00
B	20	20	36	36.00	84	84.00
C	28	28	64	64.00	64	64.00
D	22	22	86	86.00	36	36.00
E	14	14	100	100.00	14	14.00
合计	100	100.00	—	—	—	—

绘制两个地区对家电行业售后服务的质量的评价的环形图，如图 2—4：

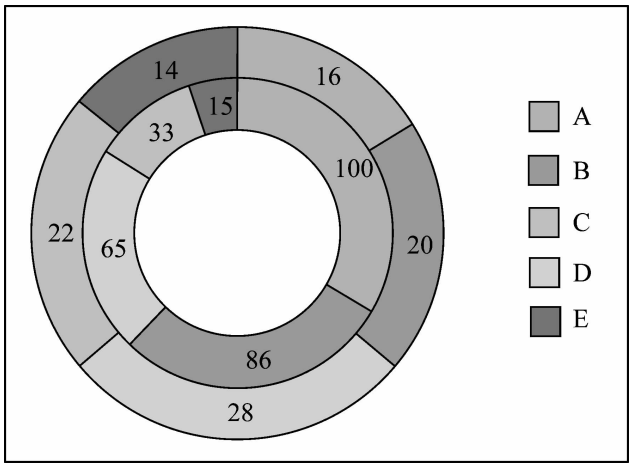

图 2—4 甲乙两地区消费者对家电行业满意程度评价

2.2.3 数值型数据的整理与显示

上一节介绍的分类数据和顺序数据的整理与显示方法，也都适用于数值型数据的整理与显示。但数值型数据还有一些特定的整理与显示方法，并不适用于分类数据和顺序数据。

1. *数据分组*

数值型数据表现为数字，因而在整理时通常是进行分组。

统计分组就是根据统计研究的目的和事物本身的特点，选择一定的标志，将研究收集的数据划分为若干性质不同的组或类的一种统计研究方法。

统计分组具有以下一些重要的性质：

首先，统计分组兼有分与合的双重功能，是分与合的对立统一。即对总体而言是“分”，要把

总体数据分为若干性质不同的部分；对个体数据而言，是“合”，要把性质相同的个体数据归入同一组中。

其次，统计分组必须遵循“穷尽原则”和“互斥原则”，即总体数据中的任何一个个体数据都必须而且只能归属于某一组，不能出现遗漏或重复出现的情况。

第三，统计分组的目的是要在同质性的基础上研究总体数据的内在差异性，即尽量体现出分组标志的组间差异而缩小其差异。因此，统计分组无论体现的是空间差异、时间差异、数量差异还是属性差异，归根到底是要客观反映各组之间本质特征的差异。

第四，统计分组在体现分组标志的组间差异的同时，却可能掩盖了其他标志的组间差异，因而任何统计分组的意义都有一定的限定性。如果需要多种角度的分组认识，就应该按多个不同的标志进行分组。

第五，统计分组的关键是分组标志的选择和分组界限的确定，如果分组标志选择不当或分组界限不合理，就会混淆事物的性质，难以客观反映现象总体的特征。当然，分组标志的选择是核心问题，因为分组界限的确定取决于分组标志。我们应该根据研究的目的，结合具体的历史条件和环境背景，选择最能体现现象本质的标志作为分组的标志。

分组后再计算出各组中数据出现的次数或频数，就形成一张频数分布表。数据分组方法有单变量值分组和组距分组两种。

(1)单变量值分组：把每一个变量值作为一组，这种分组方法通常只适合于离散变量。

表 2—6　某车间工人看管机器台数分布情况

按工人看管机器台数分组	工人数(人)	工人比重(%)
6	20	23.26
8	24	27.91
10	26	30.23
12	16	18.60
合计	86	100.00
各组变量值	次数	频率

单变量分组的注意要点：第一，将一个变量值作为一组；第二，适合于离散变量；第三，适合于变量值较少的情况。

(2)组距分组：将全部变量值依次划分为若干个区间，并将这一区间的变量值作为一组。

下面结合例子说明组距分组的过程和频数分布表的编制过程。

例 2.4　某行业管理局所属 40 个企业 2002 年的产品销售收入数据如下(单位：万元)：

152	124	129	116	100	103	92	95	127	104
105	119	114	115	87	103	118	142	135	125
117	108	105	110	107	137	120	136	117	108
97	88	123	115	119	138	112	146	113	126

试对数据进行分组。

首行考虑组距分组要点：

①将全部变量值进行排序，然后依次划分为若干个区间，并将这一区间的变量值作为一组；

②适合于连续变量，适合于变量值较多的情况；

③需要遵循“不重不漏”的原则，“不重”指一项数据只能分在其中的一组，不能在其他组中重复出现；“不漏”是指在所分的全部组别中每项数据都能分在其中的某一组，不能遗漏。

④可采用等距分组（各组的组距相等），也可采用不等距分组。

以上四条要点明确后，开始考虑分组步骤：

第 1 步：确定组数。一组数据分多少组合适呢？一般与数据本身的特点及数据的多少有关。由于分组的目的之一是为了观察数据分布的特征，因而组数多少应适中。如组数过少，数据分布就会过于集中，组数过多，数据的分布过于分散，这都不便于观察分布的特征和规律。组数的确定应以能够显示数据有分布特征和规律为目的。在实际分组时，可以按斯特奇斯提出的经验公式来确定组数 K：

$$K = 1 + \frac{\lg n}{\lg 2} \tag{2.1}$$

对结果四舍五入取整数即为组数。这只是一个经验公式，实际应用时，可根据数据的多少和特点及分析的要求，参考这一标准灵活确定组数。

将例 2.4 的数据代入式 2.1 得

$$K = 1 + 5.32 = 6.32 \approx 6$$

第 2 步：确定组距。

一个组的最大值称为上限，最小值称为下限。组距是一个组的上限与下限之差，可根据全部数据的最大值和最小值及所分的组数来确定，即

$$\text{组距} = (\text{最大值} - \text{最小值}) \div \text{组数} \tag{2.2}$$

所给数据中，最大的为 152，最小的为 87，可知数据全距为 152－87＝65。为便于计算，组距宜取 5 或 10 的倍数，而且第一组的下限应低于最小变量值，最后一组的上限应高于最大变量值。为便于计算和分析，确定将数据分为 6 组，各组组距为 10，组限以整 10 划分。

为使数据的分布满足穷尽和互斥的要求，注意到，按上面的分组方式，最小值 87 可能落在最小组之下，最大值 152 可能落在最大组之上，将最小组和最大组设计成开口组形式，即如果全部数据中的最大值和最小值与其他数据相差悬殊，为避免出现空白组或个别极端值被漏掉，第一组和最后一组可以采取“××以下”及“××以上”。开口组通常以相邻组的组距作为其组距。按照“上限不在组内”的原则，即当相邻两组的上下限重叠时，恰好等于某一组上限的变量值不算在本组内，而算在下一组内，用划记法统计各组内数据的个数——企业数，也可以用 Excel，将结果填入表内，得到频数分布表如下表中的左两列；将各组企业数除以企业总数 40，得到各组频率，填入表中第三列；在向上的数轴中标出频数的分布，由下至上逐组计算企业数的向上累计及频率的向上累计，由上至下逐组计算企业数的向下累计及频率的向下累计。

第 3 步：统计出各组的频数并整理成频数分布表。

表 2－7　40 个企业按产品销售收入分组表

按销售收入分组(万元)	企业数(个)	频率(%)	向上累计		向下累计	
			企业数	频率	企业数	频率
100 以下	5	12.5	5	12.5	40	100.0
100～110	9	12.5	14	35.0	35	87.0
110～120	12	22.5	26	65.0	26	65.0
120～130	7	17.5	33	82.0	14	35.0
130～140	4	10.0	37	92.5	7	17.5
140 以上	3	7.5	40	100.0	3	7.5
合计	40	100.0	—	—	—	—

上文提到“上限不在组内”的分组原则，对于离散变量，可以采用相邻两组组距间断的办法解决“不重”的问题；对于连续变量，可以采取相邻两组组限重叠的方法，根据“上限不在组内”的规定解决不重的问题。可以对一个组的上限值采用小数点的形式，小数点的位数根据所要求的精度具体确定。

组距分组掩盖了各组内的数据分布状况，为反映各组数据的一般水平，通常用组中值作为该组数据的一个代表值。

定义 2.7　组中值：下限与上限之间的中点值称为组中值，其计算公式为：

$$\text{组中值}=(\text{上限}+\text{下限})/2$$

组中值作为一组数据的代表值，必须有一个必要的假定条件，即各组数据在本组内呈均匀分布或在组中值两侧呈对称分布。如果实际数据的分布不符合这一假定，用组中值作为一组数据的代表值会有一定的误差。

2. *数值型数据的图示*

上一节介绍的条形图、饼图、环形图及累计分布图等都适用于显示数值型数据。此外，对数值型数据还有下面的一些图示方法，这些方法并不适用于分类和顺序数据。

通过数据分组后形成的频数分布表，可以初步看出数据分布的一些特征和规律。如果用图形来表示这一分布的结果，则更形象、直观。

(1)分组数据——直方图

直方图是用长方形的宽度和高度来表示次数分布的图形。绘制直方图时，横轴表示各组组限，纵轴表示次数(一般标在左方)和比率(或频率，一般标在右方)，若没有比率的直方图则只保留左侧的次数。依据各组的组距的宽度与次数的高度绘成直方形。根据表 2－7 的资料绘制的直方图如图 2－5。图 2－5 是依据等组距式变量数列绘制的直方图。对于不等组距式变量数列，则通常按次数密度(频数密度)绘制直方图以表示其分布。

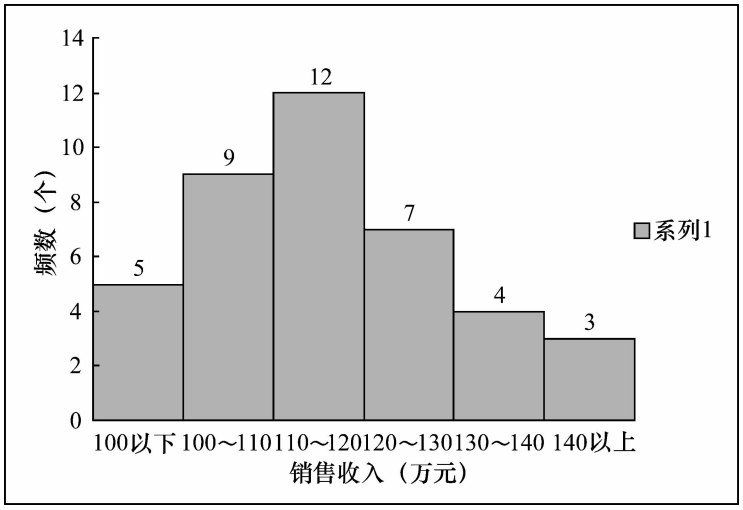

图 2－5　企业产品销售收入分布直方图

直方图与条形图的区别：

第一，条形图是用条形的高度（或长度）表示各类别频数的多少，其宽度（表示类别）则是固定的；直方图是用面积表示各组频数的多少，矩形的高度表示每一组的频数或百分比，宽度则表示各组的组距，其高度与宽度均有意义；

第二，直方图的各矩形通常是连续排列，条形图则是分开排列；

第三，条形图主要用于展示分类数据，直方图则主要用于展示数值型数据。

（2）分组数据——折线图

折线图也称频数多边形图，折线图可以在直方图的基础上，用折线将各组次数高度的坐标连接而成，也可以用组中值与次数求坐标点连接而成。图 2－6 是根据表 2－7 资料绘制的次数分布折线图。

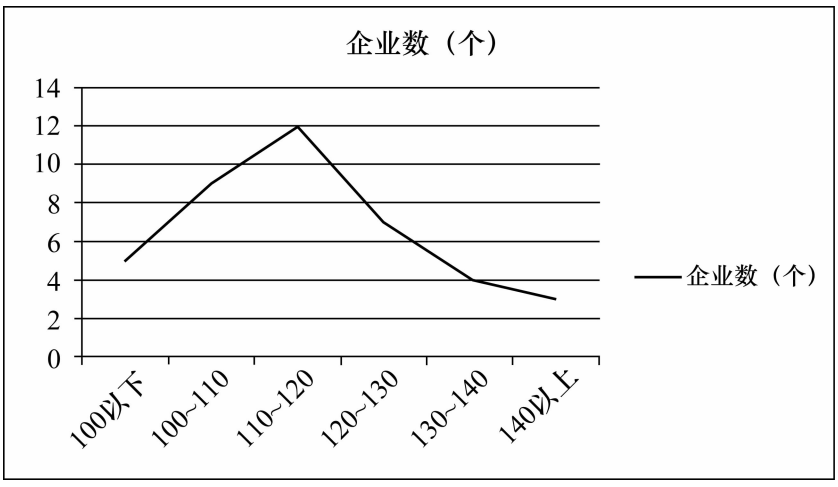

图 2－6　企业销售收入次数分布折线图

折线图的特点：