

CENTURY

21世纪普通高等院校系列规划教材

应用统计学

APPLIED STATISTICS

主 编 龙金茹
副主编 王 波 赵 元



西南财经大学出版社

中国·成都

21世纪普通高等院校系列规划教材

应用统计学

APPLIED STATISTICS

主 编 龙金茹
副主编 王 波 赵 元



西南财经大学出版社

中国·成都

图书在版编目(CIP)数据

应用统计学/龙金茹主编. —成都:西南财经大学出版社,2018. 8
ISBN 978 - 7 - 5504 - 3595 - 7

I. ①应… II. ①龙… III. ①应用统计学—教材 IV. ①C8

中国版本图书馆 CIP 数据核字(2018)第 156446 号

应用统计学

YINGYONG TONGJIXUE

主 编:龙金茹

副主编:王波 赵元

责任编辑:王青杰

封面设计:杨红鹰 张姗姗

责任印制:朱曼丽

出版发行	西南财经大学出版社(四川省成都市光华村街 55 号)
网 址	http://www.bookcj.com
电子邮件	bookcj@foxmail.com
邮政编码	610074
电 话	028 - 87353785 87352368
照 排	四川胜翔数码印务设计有限公司
印 刷	郫县犀浦印刷厂
成品尺寸	185mm × 260mm
印 张	17.25
字 数	370 千字
版 次	2018 年 8 月第 1 版
印 次	2018 年 8 月第 1 次印刷
印 数	1—2000 册
书 号	ISBN 978 - 7 - 5504 - 3595 - 7
定 价	38.60 元

1. 版权所有,翻印必究。
2. 如有印刷、装订等差错,可向本社营销部调换。
3. 本书封底无本社数码防伪标识,不得销售。



Foreword 前言

世界正步入大数据时代，几乎所有事物都可被量化。然而数据只是最原始的形式，只有通过整理、分析才有意义。统计学就是破译数据内在联系的工具。

统计学是一门关于数据的科学，是关于数据收集和数据分析的方法论科学。统计学因此也成为经济类和工商管理类专业的核心基础课程。具备了统计学的知识，我们就可以在数据的海洋中遨游，并发现数字背后的真相和规律。

本书具有下列特点：

(1) 本书强化应用和动手操作。本书致力于将统计学的应用作为出发点和归宿，开辟独立的篇章和大量篇幅进行 EXCEL 常用操作详解，将统计的理论和应用衔接起来。本书对高级筛选、排序、分类汇总、数据透视表、直方图、柱形图、条形图、饼图、折线图、面积图等常用统计图表的制作和应用进行了详细阐述；对描述统计中集中趋势和离散程度的各类指标用 EXCEL 公式来实现；对推断统计中的参数估计、假设检验、线性回归分析也用 EXCEL 来实现，并将实现的详细步骤写入书中，力图培养学生的学习兴趣、提高学生的动手能力。

(2) 本书深入浅出、通俗易懂，主要强化基础知识，避免复杂的数学公式推导。本书采用逻辑叙述和分析的方法，通过易懂的语言配以卡通图片、图表，解释公式背后的统计思想，避免枯燥的数学公式推导。

(3) 本书较系统、较全面。内容从基础概念、数据处理、描述统计到推断与分析综合，涵盖了统计学的重要章节和内容，并为每章配备了一定的练习思考题目。

全书由龙金茹担任主编，昆明理工大学的王波和国家计算机网络与信息安全管理中心云南分中心的赵元担任副主编。其中，龙金茹负责全书大纲的设计，并负责第一章、第二章、第三章、第四章、第五章、第六章以及 EXCEL 模块 1 至模块 3 部分；王波负责第七章、第八章以及相应章节 EXCEL 操作的编写；赵元负责 EXCEL 的模块 4 至模块 6 以及习题练习的编写。龙金茹对全书进行了最后的修改和总纂。

本书在编写过程中引用了有关专家和学者的研究成果和资料，在此深表谢意。同时，感谢西南财经大学出版社员工为出版本书所付出的辛勤劳动。

由于编者水平有限，加之编写时间仓促，书中难免有疏漏之处，恳请各位读者批评指正！

编者

2018 年 7 月



Contents

目录

1 绪论	1
1.1 统计的含义	1
1.2 统计研究的特点	3
1.3 统计学的分支	3
1.4 统计数据类型	5
1.5 统计学的基本概念	7
1.6 统计学的发展历程	11
习题	13
2 数据整理与显示	16
2.1 统计数据的审核	16
2.2 统计数据的分组	17
2.3 频数分配	22
2.4 频数表的图示法	24
2.5 其他常用的统计图	25
习题	30
3 数据的概括性度量	33
3.1 总规模度量——总量指标	34

3.2	比较度量——相对指标	35
3.3	集中趋势的度量——平均指标	39
3.4	离散程度的测度——变异指标	53
3.5	偏态和峰度的测度	59
	习题	62
4	抽样推断与分布	65
4.1	抽样推断的概述	66
4.2	抽样方法	67
4.3	抽样分布	71
4.4	抽样误差	77
	习题	79
5	参数估计	81
5.1	参数估计的基本问题	81
5.2	总体均值的区间估计	87
5.3	总体方差的区间估计	90
5.4	两个总体方差比的区间估计	91
5.5	样本容量的确定	92
	习题	94
6	假设检验	97
6.1	假设检验的基本原理	97
6.2	假设检验的流程	105
6.3	总体均值的假设检验	106
6.4	总体方差的假设检验	109
6.5	两个总体方差比的假设检验	111
	习题	113

7 相关与回归分析	115
7.1 变量名称及其关系	116
7.2 相关关系的描述与度量	117
7.3 一元线性回归	120
7.4 多元回归方程	129
习题	136
8 时间序列分析	139
8.1 时间序列分类及编制原则	139
8.2 时间序列的水平分析	141
8.3 时间序列的速度分析	145
8.4 时间序列的趋势分析	147
8.5 时间序列的季节变动分析	154
习题	159
9 Excel 2010 操作	161
模块一 数据的预处理	161
模块二 统计图	173
模块三 直方图	202
模块四 概括性的数字	211
模块五 参数估计	228
模块六 假设检验	241
模块七 相关与回归分析	254
模块八 时间序列	260
参考文献	270



1 绪论



学习目标

1. 理解“统计”的含义、统计研究的特点。
2. 掌握统计学的分支、数据类型、统计学中的几组基本术语。
3. 能基于数据指出数据的类型。
4. 能结合实际中遇到的具体事例说明总体、样本、变量、变量值、样本个数和样本容量。
5. 了解统计学的发展过程、统计学的应用范围。

1.1 统计的含义

统计是人类认识社会的一种有力武器，在各种实践活动和科学研究领域经常使用。“统计”一词有三种含义，即统计活动、统计资料和统计科学。

(1) 统计活动，又称统计工作，是指收集、整理和分析统计数据，并探索数据内在的数量规律性的活动过程。比如，球类比赛时解说员要统计竞赛双方的进攻次数和成功率。

(2) 统计资料，或称统计数据，即统计活动过程所获得的各种数字资料和其他资料的总称，如统计年鉴、统计分析报告、政府统计公报等。

(3) 统计科学，即统计学，是阐述统计工作基本理论和基本方法的科学，是对统计工作实践的理论概括和经验总结。

统计学究竟是什么，至今没有一个统一的定义。《大不列颠百科全书》对统计学的定义为：统计学是收集、分析、表述和解释数据的科学。《中国百科全书·数学卷》对统计学的定义为：统计学是一门科学，它研究怎样以有效的方式收集、整理、分析带随机性的数据，并在此基础上对所研究的问题做出统计学推断，直至对可做出的决策提供依据或建议。

综合来说，统计学是一门收集、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。由于统计学是与数据打交道的科学，因此，统计学也被称为“数据的科学”。

统计的三种含义不是孤立存在的，而是有着非常密切联系的。统计工作是统计的最基本含义，是统计资料和统计学的前提和基础，统计资料是统计工作的成果，统计学是统计工作的经验总结和理论升华。对统计学的深入研究又会大大提高统计工作和统计资料的质量。三者的关系如图 1-1 所示。统计活动、统计资料、统计科学相互依存、相互联系，共同构成了一个整体，这就是通常所说的统计。

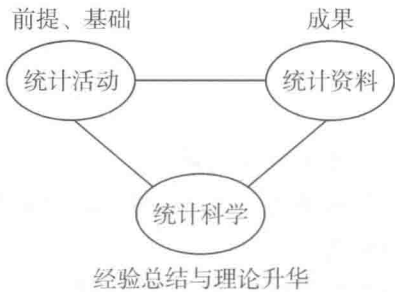


图 1-1 统计三种含义的关系

那么，什么是数据内在的数量规律性呢？下面用两个简单的例子来说明。

【例子 1-1】我们都知道，随机地投掷一次硬币不能事先确定出正面或反面，但经过大量观察，不断重复试验时，就会发现一枚硬币出现正面和反面的概率会大体相同，即 $0.5 : 0.5$ 。这也就是我们所探索的数量规律性。历史上有多位科学家探索过这一规律性。18 世纪法国博物学家布冯伯爵抛一枚硬币 4 040 次，出现 2 048 个正面，正面出现的概率为 0.506 9；数学家科里奇在二战期间被德国人关进监牢期间，抛硬币 10 000 次，出现正面的次数 5 067 个，正面出现的概率为 0.506 7；1900 年前后，英国生物统计学家皮尔逊抛一枚硬币 24 000 次，正面出现的次数为 12 012 个，正面出现的概率为 0.500 5。试验次数越多，出现正面和反面的可能性就越接近 $1/2$ 这一稳定的数值。

【例子 1-2】一个家庭新生婴儿的性别可能是男也可能是女，从表面上看，新生婴儿的性别比例似乎没有什么规律可循。但如果对新生婴儿的性别进行大量观察，就会发现性别比例还是有规律的，即婴儿总人数中男孩要多于女孩，大致为每生育 100 个女孩，就有 105 个左右的男孩。这个 $105 : 100$ 就是新生婴儿男女性别的数量

规律，它是人类社会长期遗传与发展的结果。有人会问，新生儿男多于女，不是性别不平衡了吗？是的，新生儿男多于女会出现不平衡，但是男婴夭折率高于女婴，到了中青年时，男女人数就大致相同了。进入中老年后，男性的死亡率仍然高于女性，导致男性的平均预期寿命比女性短，长寿的男性要少于女性，中青年男女人数大致相同，老年时女性又略多于男性。这一现象既保证了人类在中青年结婚生育时性别大致平衡，又使得在人口总数上男女也大体相当，有利于人类社会的进化和发展。对人类性别比例的研究是统计学的起源之一，也是统计方法探索的最早的数量规律之一。

1.2 统计研究的特点

统计研究的特点可以归纳为以下几个方面：

(1) 数量性。统计学的特点是用大量数据资料说明事物的规模、水平、结构、比例关系、发展速度等。比如例子 1-1 中硬币出现的正面次数与反面次数之比为 1:1，例子 1-2 中新生儿性别比的数量关系为 105:100。

(2) 总体性。个别现象的特征往往具有特殊性、偶然性，只有现象总体的特征才具有相对的普遍性、稳定性，所以研究总体的特征才能科学、准确地认识社会经济现象。以总体为研究对象，才能排除偶然性因素的影响，从而揭示出事物的内在联系和规律性。统计研究关心的是总体，但是研究过程是从个体出发，对大量个体进行登记、整理，进而过渡到总体的数量方面。

(3) 具体性。统计研究的数量方面指社会经济现象的具体的数量方面，而不是抽象的纯数量关系，这是它不同于数学的重要特点。统计研究的数据具有时间和空间的规定性。如果抽掉具体的时间、空间和条件进行研究，数字就会显得空洞无物，也就不是统计数字了。如，2008 年全年进出口总额 25 616 亿美元，比上年增长 17.8%。

(4) 社会性。社会经济现象的产生源于人们对社会活动的参与，并且我们所需要的资料源于并服务于全社会，同时资料的取得与维护需要全社会的积极配合与参与。可见，统计的研究对象具有明显的社会性。

1.3 统计学的分支

统计学的内容十分丰富，研究与应用的领域非常广泛。从统计教育的角度，统计学大致有以下两种分类：

1. 描述统计和推断统计

描述统计指用表、图、概括性的数字对数据进行描述的统计方法。描述统计是对数据进行处理的第一阶段，用图、表和概括性数字（如平均数）表示数据的分布、形状特征，并为进一步的统计推断提供依据。推断统计指根据样本信息对总体进行估计、假设检验、预测或其他推断的统计方法。如，想要调查影响消费者网络购物的因素，随机抽取了1 000名网购消费者，调查什么因素会影响他们最终的网购结果，发现商品价格、详情页的精美度以及售后质量是最主要的原因，并据此推断出，消费者网购的重要影响因素就是价格、商品详情页以及售后服务。此种分类一方面反映了统计发展的前后两个阶段，另一方面也反映了统计方法研究和探索客观事物内在数量规律性的先后两个过程。图1-2即为统计方法探索客观现象数量规律性的过程图。

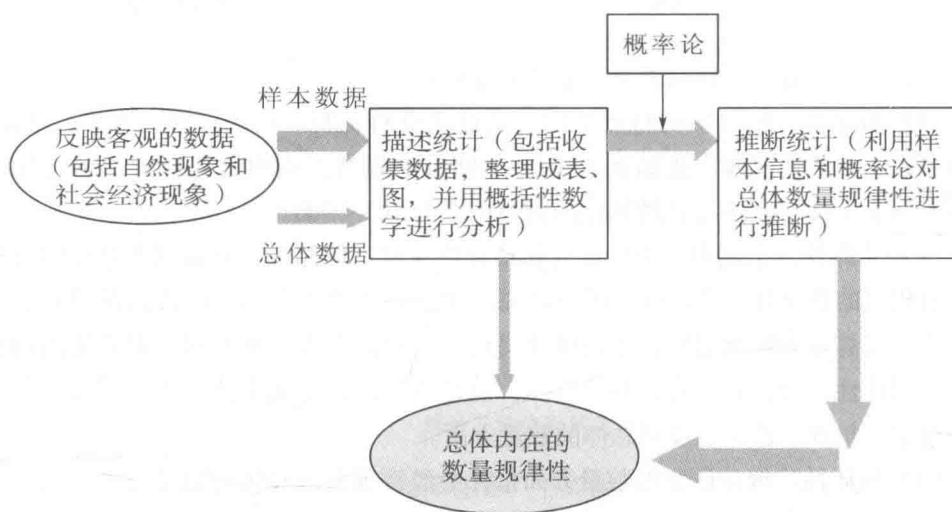


图1-2 统计学探索客观现象数量规律性的过程图

2. 理论统计和应用统计

理论统计是指以概率论为基础，从纯理论的角度，对统计方法加以推导论证，中心的内容是统计推断问题，如参数估计、假设检验、方差分析、相关分析、回归分析等。应用统计是指以各个不同领域的具体数量为研究对象，运用理论统计学中的方法解决具体领域的问题，如生物医学统计、国民经济统计。生物医学统计指用理论统计的原理和方法，分析和解释生物界的种种现象和数据资料，以求把握其本质和规律性。

在统计工作者中，从事理论统计研究的人只是很少的一部分，大部分是应用统计的方法去解决实际问题的应用统计工作者。统计方法扩展到了几乎所有的研究领域。不论是物理统计、医学统计还是管理统计，其统计的主要方法都是描述统计和推断统计。

作为一名优秀的应用统计工作者，不但要熟练掌握和应用各种统计方法，而且必须具备所研究和应用领域的专业知识。因此，专业知识和统计方法是做好统计应用的两项基本功。例如，要做好经济统计工作，其基本条件是要具备经济理论和统计方法的知识能力，当然还要有计算机、写作和其他能力。这就要求统计应用人才必须是复合型人才，这也是人文、社科、理、工、农、医等院校都开设统计方法这门课程的缘由。

1.4 统计数据类型

统计数据是对现象进行测量所得到的结果。比如，对班级实训课的五档制成绩进行测度，得到优、良、中、及格、不及格的成绩数据；对股票价格变动水平的测量可以得到股票价格指数数据；对某省利用外商直接投资情况进行测度，可以得到 FDI 数据；对人口性别的测量可以得到男或女这样的数据。下面从不同角度说明统计数据的分类。

1.4.1 分类数据、顺序数据、数值型数据

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

1. 分类数据

分类数据是指只能归于某一类别的文字型数据。它是对事物进行分类的结果，数据表现为类别。例如，人口按照性别分为男、女两类；企业按照所有制性质分为私有企业、国有企业、外商投资企业；人口按照民族分为汉族、壮族、回族等。这些都属于分类数据。

分类数据没有优劣之分，支持的运算符号为等号“=”、不等号“ $\neq$ ”。例如，性别的取值有男性和女性。“男性”和“女性”之间没有优劣之别，他们是平等的，因此支持的运算符号要么是“男性 $\neq$ 女性”“男性=男性”或者“女性=女性”。

2. 顺序数据

顺序数据是指只能归于某一有序类别的非数字型数据。顺序数据虽然也是类别，但是这些类别是可以排序的，即有优劣之分，但是它们之间具体的差异是无法计算的。比如产品分为一等品、二等品、三等品、次品等；考试成绩分为优、良、中、及格、不及格等；受教育的程度分为小学、初中、高中、大学及以上；对某事物的评价可以分为非常满意、满意、一般、不满意、非常不满意等。我们拿考试成绩来说，“优”比“良”好，只是“优”比“良”具体优异多少是无法计算的；满意程度而言，“非常满意”肯定比“一般”要好，只是具体数值无法进行计算。因此，

顺序尺度的数据支持的运算符号为 $<$ 、 $>$ 、 $\leq$ 、 $\geq$ 以及 $=$ 和 $\neq$ 。此类运算是成立的，如优 $>$ 及格、不满意 $<$ 非常满意、初中 $\neq$ 高中。

3. 数值型数据

它是按照数字尺度测量的观察值，其结果表现为具体的数值。现实中处理的大多数数据都是数值型数据。数值型数据支持所有的运算符号，如 $\leq$ 、 $\geq$ 、 $\neq$ 、 $+$ 、 $-$ 、 $\times$ 、 $\div$ 、幂、指数运算等。国内生产总值（GDP）、身高（cm）、体重（kg）、股票价格（元）、销售量（台）就是数值型数据。

分类数据和顺序数据说明的是事物的品质特征，通常都是用文字进行表述，因而也可以统称为品质数据或定性数据；数值型数据说明的是现象的数量特征，通常都是用数值来表述，因此也可以称为定量数据或数量数据。三者的具体特点及支持的运算见图 1-3。

	品质数据、定性数据		数量数据、定量数据
	分类数据	顺序数据	数值型数据
特点	数据表现为文字； 数据无优劣之分	数据表现为文字； 数据有优劣之分，可以 排序，但是无法计算具 体差异	数据表现为数值； 数据有优劣之分，并 可以计算具体差异
支持的 运算	$=$ 、 $\neq$	$=$ 、 $\neq$ $\geq$ 、 $\leq$ 、 $>$ 、 $<$	$=$ 、 $\neq$ 、 $\geq$ 、 $\leq$ 、 $>$ 、 $<$ 、 $+$ 、 $-$ 、 $\times$ 、 $\div$ 、幂、指数

图 1-3 分类数据、顺序数据和数值型数据的图解

1.4.2 截面数据、时间序列数据、面板数据

按照被描述的现象与时间的关系，可以将统计数据分为截面数据、时间序列数据和面板数据。截面数据是在相同或近似时间点上收集的数据，这类数据通常是在不同的空间或个体上获得的，用于描述现象在某一时刻的变化情况。比如，2017 年我国各地区的地区生产总值数据、某日有色金属板块中的股票收盘价数据、本学期某课程学生的考试成绩数据，这些都是截面数据。时间序列数据是指在不同的时间收集到的相同空间或个体的数据，这类数据是按时间顺序收集到的，用于所描述现象随时间变化的情况。比如 1999—2017 年我国的国内生产总值数据、江西铜业股票 1—5 月的每日收盘价、某学生四个学期以来的综合测评成绩数据，都是时间序列数据。面板数据也称为混合数据，是在不同的时间、不同的空间或个体上获得的数据。

比如 2000—2017 年我国各地区的地区生产总值数据、1—5 月有色金属行业 50 只股票的每日收盘价数据、四个学期以来班上 50 名学生的期末总评成绩数据，都是面板数据。

1.4.3 观测数据、实验数据

按照统计数据的收集方法，可以将其分为观测数据和实验数据。观测数据是通过调查或观测而收集到的数据，这类数据是在没有对事物进行人为控制的条件下得到的，有关社会经济现象的统计数据几乎都是观测数据。观测数据是无法重复获得的，即“历史不会重演”。实验数据则是在实验中控制实验对象而收集到的数据，实验数据可以重复获得。比如，对一种新药疗效的实验数据，对一种新的农作物品质的实验数据。自然科学领域的大多数数据都是实验数据。

区分数据类型很重要，因为对不同类型的数据，需要采用不同的统计方法来处理和分析。比如，对分类数据，我们通常计算各组的频数或频率、众数、异众比率等，进行列联表分析和 χ^2 检验等；对顺序数据，可以计算中位数、等级相关系数等；对数值型数据，可以用更多的统计方法进行分析，如参数估计、假设检验等。统计数据的分类见图 1-4。

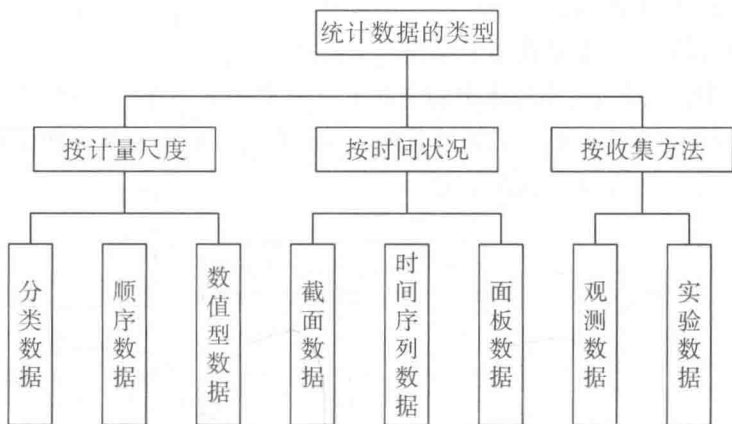


图 1-4 统计数据的分类

1.5 统计学的基本概念

统计学中的概念很多，其中有几个概念是经常要用到的，有必要单独加以介绍。这些概念包括总体和总体单位、样本和样本单位、变量和变量值、样本个数和样本容量。

1.5.1 总体、总体单位

总体是指在一定统计研究目的下的所有调查对象构成的全体（示例见图 1-5）。它是由客观存在的、具有某种共同性质的许多个体组成的整体。构成总体的这些个别单位则称为总体单位。例如，要研究某市非公有制工业企业的生产经营情况，那么该市全部非公有制工业企业就构成一个总体，因为它们是需要调查的对象，只有从它们身上才能调查出企业的生产经营情况如规模、利润、员工人数等，该市每一家非公有制工业企业就是总体单位。

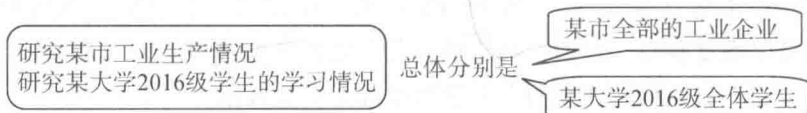


图 1-5 一定研究目的下的统计总体示例

需要注意的是，在统计研究过程中，统计研究的目的和任务居于支配和主导地位，是考虑问题的出发点。统计总体取决于统计研究的目的和任务，有什么样的研究目的就要求有什么样的统计总体与之相适应，随后的统计调查、统计整理和统计分析都是围绕总体来进行。

总体具有同质性、差异性和大量性三个特点。

（1）同质性。同质性指总体中的全部单位至少在某一方面具有共同的特征。各个个体之所以可以放在一起形成统计总体，就是因为它们有着某一方面相同的性质。同质性是构成总体的基础。如图 1-6 所示。

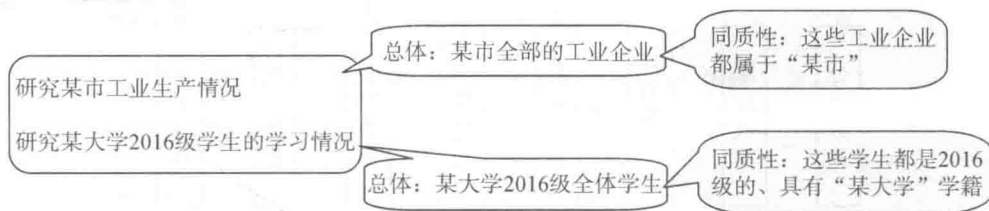


图 1-6 统计总体同质性的体现

（2）差异性。差异性指总体中的各个单位除了至少某一方面具有共同的特征之外，在其他方面存在差异性。如图 1-6 中所示，某市全部的工业企业这个总体中，除了具有属于“某市”这个共同点之外，这些企业在资产规模、主营业务销售额、资产负债率、每股收益率等方面都是不同的；某大学 2016 级全体学生这个总体中，除了同质性之外，这些学生在身高、民族、体重、经济状况、成绩等方面都是不同的。总体更多地是体现差异性。如果总体单位不存在差异性，也就是说，各个总体单位都是同质的，那么只要抽出其中一个单位进行研究，就能代表总体，从这个意义上说，统计研究就没有必要了。如果某市工业企业没有差异性，那么只需要抽其

中一家企业了解情况就能知道总体，统计研究也就没有必要了。因此，总体的差异性是统计研究的基础。

(3) 大量性。大量性是指相对于任务而言具有足够多数量的同质性个体（示例见图 1-7）。只有一个单位的统计总体是不存在的。当然，根据研究目的的不同，总体也不一样，总体中所包含的总体单位的数量也就不同，一个总体究竟包含多少总体单位，最终取决于统计研究的目的。



图 1-7 统计总体大量性的示例

1.5.2 样本、样本单位

统计研究的目的是确定总体的数量特征，但是有时总体的单位数太多或无限，不可能也没有必要对每个总体单位做调查，这时，就要借助样本研究总体了。样本是按照一定概率从总体中抽取并作为总体代表的一部分总体单位的集合体。

样本是统计学中非常重要的概念，对这一概念的理解要注意三个方面的问题：第一，构成某一样本的每一个样本单位必须取自某一特定总体，不允许总体之外的单位介入；第二，样本单位的抽取是按照一定概率进行的，具体样本的产生应该是随机的，必须排除人为主观因素的干扰；第三，样本是总体的一部分，带有总体的信息，因而能够推断总体，但是，样本毕竟只是总体的一个子集，且具有随机性，故由样本去推断总体会产生误差。

1.5.3 变量、变量值

表示现象某种特征或者属性的概念称为变量。变量的具体表现称为变量值。例如，“性别”是一个变量，其变量值是“男”或“女”；“身高”也是一个变量，其变量值可以是 155 厘米，178 厘米，190 厘米，…；“成绩”是一个变量，变量值可以是“优”“良”“中”等。

按照变量值的取值是否连续，变量分为连续型变量和离散型变量两种。连续变量指变量值是连续不间断的变量，其变量值可以取整数也可以取小数，变量值不能一一列举，如“零件尺寸”“身高”“资金”“利润”等。离散型变量指变量值之间