柴根象　钱伟民　编著

统计学教程

同济大学出版社

统 计 学 教 程

柴根象　钱伟民　编著

同济大学出版社

内 容 提 要

本书主要介绍统计学的基本概念、基本原理和基本方法，强调直观性，突出统计思想的阐述。全书由数据的整理和描述、抽样推断和抽样分布、参数点估计、区间估计、假设检验、回归分析和方差分析等6章内容组成。书中附有相当数量的习题。

本书可供统计学专业、应用数学专业和其他理、工科专业的本科生作为教材，也可供工科硕士研究生和自学者参考。

图书在版编目(CIP)数据

统计学教程/柴根象，钱伟民编著．—上海：同济大学出版社，2004.1

ISBN 7-5608-2625-3

Ⅰ．统… Ⅱ．柴… Ⅲ．统计学—高等学校—教材 Ⅳ．C8

中国版本图书馆CIP数据核字(2003)第110856号

统计学教程

柴根象　钱伟民　编著

责任编辑　李炳钊　责任校对　郁　峰　封面设计　李志云

出　版 发　行	同济大学出版社
	(上海四平路1239号　邮编200092　电话021-65985622)
经　销	全国各地新华书店
印　刷	大丰印刷二厂印刷
开　本	787mm×960mm　1/16
印　张	15.75
字　数	315000
印　数	1—2100
版　次	2004年1月第1版　　2004年1月第1次印刷
书　号	ISBN 7-5608-2625-3/O·248
定　价	20.00元

前　言

统计学在我国高校的绝大部分理、工科专业以及管理类专业中都已成为一门重要的基础课程。这不仅因为统计学在自然科学和社会科学的各个领域得到了广泛的应用，而且，从人才素质的全面培养来说，这门课程也是不可或缺的。进入21世纪以来，随着网络技术的发展，人们可以更直接、更便捷地获得越来越多的统计信息，这些统计信息传递着各种市场的信息和政府部门的重要政策取向。没有良好的统计知识，就不可能很好地把握这些统计信息的特性并合理加以运用。

本书着眼于介绍统计学的基本概念、基本原理和基本方法，强调直观性，突出统计思想是本书的特点。本书的大部分内容适用于理、工科的本科生教学，也有部分内容适用于硕士研究生教学，如第二章的第六节，第四章的第四节（目录中打＊号部分），第五章的第六节，第六章的第三、四节等。

本书第一稿的前、后两部分分别由柴根象、钱伟民撰写，并形成一个讲义，在同济大学相关专业试用了两年。在讲义的基础上，由钱伟民撰写本书的第二稿。定稿时，由柴根象对第二稿作了适当的修改。杨筱菡同志协助完成了本书的习题解答。

本书的出版得到同济大学教材出版基金的资助。同济大学出版社给予了大力支持，在此我们表示真诚的谢意。

由于作者学识有限，书中错漏及不当之处在所难免，敬请各位同行及读者不吝赐教，以便日后修订时加以改进。

编　者

2003年8月

目　录

第一章　数据的整理和描述

随着信息时代的到来和计算机网络技术的发展，人们越来越多地接触到大量的统计数据．如何在大量杂乱无章的数据中“挖掘”出有用的信息，是放在我们面前的一个重要课题．统计学提供了整理和分析数据的有用的方法．本章介绍描述性统计学的基本内容，通过对数据进行整理和描述方法的学习，对统计学有一个直观的了解．

第一节　统计数据和统计数据的收集

1.1.1　数据的类型

统计数据是统计学研究的出发点，也是统计加以实施的载体，因此有必要研究统计数据的分类、结构和特点.

先看一些例子：

例 1.1　某公司拥有多个有关其雇员、客户和商务经营的数据库. 这些数据库中有雇员的年龄、性别、文化程度、薪水和服务年限等个人记录，也有关于销售额、广告支出、发售成本、库存水平和生产数量的详细数据，还有关于客户情况的详细资料，这些数据构成了公司的内部信息资源.

例 1.2　某医院为在某地区比较吸烟与肺癌之间有无关联，在该地区按一定的方法随机地抽取一批人，按是否吸烟分成两组进行对照试验，得到吸烟患肺癌与不吸烟患肺癌的比率.

例 1.3　证券市场每天都会产生大量的证券交易数据，这些数据通过媒体传播到世界各地. 表 1.1 给出了三个银行股的四天的收盘价(单位:元).

表 1.1　　**股票价格**　　**(单位:元)**

代　码	时间 / 股票名称	2002 年 9 月 2 日	2002 年 11 月 29 日	2002 年 12 月 26 日	2003 年 1 月 2 日
600000	浦发银行	12.52	10.37	9.62	9.25
600016	民生银行	12.89	10.94	9.83	8.90
600036	招商银行	10.79	9.21	8.47	8.06

例 1.4 表 1.2 给出了北京市总人口从 1982 年到 1990 年的数据.

表 1.2　　1982—1990 年北京市总人口　　(单位:万人)

年份	1982	1983	1984	1985	1986	1987	1988	1989	1990
总人口	917.82	933.20	945.20	957.90	971.23	987.97	1001.20	1021.11	1032.21

根据在描述事物时所用的度量尺度的不同,数据可分为**数量数据**和**品质数据**. **数量数据**用来说明事物的数量特征. 例如,股票的收盘价、北京市今年的总人口数等都是数量数据. **品质数据**又称为**分类数据**,它是用来说明事物的品质特征. 例如,公司雇员的性别、文化程度、吸烟与否、患肺癌与否等,都可用品质数据来描述. 品质数据可用文字形式来表现,但有时为了方便,也用数量来表示. 例如,不同的股票名称是品质数据,但我们也常用股票代码来表示股票名称.

数据按照被描述的对象与时间的关系分为**截面数据**、**时间序列数据**和**平行数据**.

截面数据是在相同或近似相同的时间点上收集的数据,用于描述事物在某一时刻的变化情况. 例如,报刊上登载的前一日的股市行情和汇市行情数据,高血压病人在服用某降压药一周后的血压值等.

时间序列数据是在一定的时间范围内收集到的数据,用于描述事物在一定的时间范围内的变化情况. 例如,例 1.4 中给出的 1982—1990 年北京市总人口的变化情况.

平行数据是截面数据和时间序列数据的组合,用于描述几个对象在一定的时间范围内的变化情况. 例如,例 1.3 中给出的三个银行股在一段时间内收盘价的变化情况.

1.1.2 统计数据的结构

一般说来,统计数据含有所描述的对象的信息,因而可用于对描述的对象的推断. 例如,在例 1.1 中根据从公司中随机选取到 100 位员工的年龄,可以算出其算术平均数,这个数值可以用来估计这个公司员工的平均年龄. 然而,各种统计数据含有的被研究对象的信息是有很大差别的. 如果获取数据的方法不当,就有可能得不到所研究的对象的信息,甚至提供错误的信息. 例如,在例 1.2 中,如果在吸烟组中都抽取中老年人,而在不吸烟组中大多选取青年人乃至少年儿童,那么,不吸烟组的患肺癌比率自然偏低.

统计数据的误差有两种,一是系统误差,二是随机误差. **随机误差**是由选取调查对象的随机性引起的误差. 例如,在例 1.2 中,为研究吸烟组患肺癌的比率,可以用相同的选取方式从某地区选取两组(每组各 100 位吸烟者)进行调查,我们可以发现,这两组人中患肺癌的比率也不完全一样,当选取方法合理时,这种差异就是由随机误差

引起的.

系统误差和随机误差相比有很大的差异. **系统误差**通常由人为的过失引起,它以相同的方式影响所有数据,将它们推向同一方向. 例如,在例 1.2 中,如果由于试验安排不当,在不吸烟组中大多选取青少年,患肺癌比率就明显偏低,如果抽取 m 组(每组 100 个不吸烟者),由此得到各组的患肺癌比率 $p_1,\cdots,p_m$ 明显偏小. 随机误差的影响则是随着不同次的观察而变化的,将统计数据推向不同的方向,有时偏大,有时偏小. 由此可见,系统误差的存在使得我们在推断时会造成严重的偏差,因而必须尽量避免系统误差.

从上面的讨论可以得出,统计数据在结构上有如下的模式:

单个观察数据＝真值＋偏差＋随机误差

其中,偏差是由系统误差引起的. 我们注意到统计数据都有随机误差,因此,随机误差这一项是不能消除的. 如果随机误差的影响过大,则会影响统计推断的精度,因此,如何降低或者控制随机误差造成的影响是统计方法研究所追求的一个目标. 偏差项存在与否以及偏差的大小是衡量统计数据质量的一个重要方面,因此,在数据收集阶段,要注意尽量消除偏差.

1.1.3　数据的收集

从上节讨论知道,偏差项存在与否可以作为衡量统计数据质量的一个重要方面. 而偏差又是由系统误差引起的,因此,要得到高质量的统计数据,就要避免系统误差,在此前提下,减少随机误差的影响. 要做到这些,就要研究如何有效地收集数据.

收集数据是统计学的一项基础性工作,收集的数据质量好坏对后续工作有重大影响. 如果数据有严重的偏差,或虽无偏差但随机误差的影响过大,则纵然使用良好的推断方法,也得不出好的结果.

统计学中有两个分支研究收集数据的统计方法:一是试验设计;二是抽样调查. 本节简明扼要地介绍这两个分支的研究内容,其中的细节可参阅相关书籍.

(1) 试验设计

试验设计顾名思义就是试验、观察的科学安排. 试验设计应该至少满足如下两个条件:1. 随机性,即完全客观地收集试验对象,并按随机方式安排其进入试验或观察,其要点是排除人为的干扰,例如,医学试验中常用的随机双盲设计就满足这一要求,随机双盲设计要求在评价某种药物的疗效时医生和病人均不知道其选用的是药物还是安慰剂,只有试验的安排者知道具体的情况,这样就使医生和病人对药物的评价客观、公正. 2. 尽可能地排除混杂因素. 例如,在例 1.2 中,如果安排 10 个吸烟者作为处理组,另 10 个不吸烟者作为对照组,相继观察若干年获得两组中患肺癌人数

的数据．这样的设计过于粗糙，因为其中至少含有两个混杂因素：年龄和性别．事实上，老年和青少年患肺癌的比率是显著不同的；同样，男性和女性患肺癌的比率也有较大差别．一个合理的设计应将同年龄段同性别的试验者分成吸烟与不吸烟的两组进行观察，这样的设计就排除了年龄和性别这两个混杂因素，或者说混杂因素得到了控制．

完全随机化设计是一种最为简单的试验设计，它完全用随机化方式决定试验单元如何分配到要比较的项目中去．如在例 1.2 中，可先依随机方式征集符合条件的志愿者（如规定的年龄段、性别及基本健康状况的志愿者）．例如征集到 500 人，其中吸烟与不吸烟的人各占一半，然后分别在 250 个吸烟者与 250 个不吸烟者中以随机方式各产生 25 组，每组 10 人．最后从 25 组吸烟者中随机抽出一组与从 25 组不吸烟者中随机抽出的一组搭配成为一个对照试验组合，这样一共产生 25 个对照试验组合．在这个设计中，志愿者是在该地区满足条件的人群中随机产生的；其次，在 250 个吸烟志愿者及 250 个不吸烟的志愿者也是以随机方式分配到 25 个对照试验组合中去的，而且在同一对照试验组合中，除去吸烟与否外，其他方面的条件大致相同．

在完全随机化设计中，对随机化不加任何限制，其优点是能消除数据的偏差；但另一方面，有可能加大了随机误差的影响．由此，在完全随机化设计的基础上又可派生出许多其他的设计方案．这里不一一列举了．

(2) 抽样调查

抽样调查在社会、经济领域有着广泛的应用．直观上看，只要在研究对象中抽取足够多的试验对象，通过试验或观察得到足够的数据，这些数据能够提供的研究对象的信息也就会有很多，这一般说来是对的，但也有例外，如果设计的抽样方案有重大缺陷，即使获取了足够多的数据，也是于事无补的，它只能在较大规模下去重复基本的错误．抽样方案的设计原则与试验设计大致相同，即避免出现系统误差，同时抽样方案的设计也有它的特点，它必须注意调查方法，特别是当调查项目涉及某些敏感问题或与被调查人有利害关系时，如不注意调查方法，很可能得不到真实的数据，从而使数据存在较大的偏差．

抽样调查方案有很多种，最为简单易行的是**简单随机抽样**，其要旨是使研究对象中每个个体以同等机会被抽到．简单随机抽样的实施十分方便，只要使用一张随机数表就可以了．

当研究对象规模很大且各个个体很分散时，使用简单随机抽样费用昂贵或者很难实施，此时多使用**集团抽样方案**或称为**整群抽样方案**，即将研究对象中一些相近个体组成集团，然后将每一集团视为一个个体，再用简单随机抽样方法抽取一个或若干个集团，试验或观察这若干个集团中的每个个体得到数据，由于在整个方案实施中地域更集中了，这样得到的数据所需的费用较低，但由于在某种程度上牺牲了“机会均

等”原则，抽出的个体的代表性要差一点．

还有其他的抽样调查方案，如分层抽样、系统抽样方案等，限于篇幅，不再一一介绍，有兴趣的读者可参阅书目[12].

上述用随机化方式设计的试验或抽样方案，其目的是用偶然的自发作用去控制和消除系统偏差，但决不是说，这种自发作用愈大愈好，随机误差的影响太大，也会影响推断的精度．一般来说，在设计试验或抽样方案时，应在系统偏差得到控制的前提下，尽量减少随机误差的影响．

第二节　描述统计学Ⅰ：表格法和图形法

从抽样调查或设计的试验中所得到的数据通常是十分庞杂的，必须经过整理和归纳才能显示出这批数据所遵循的规律和含有的有关研究对象的信息．整理数据的常用且直观的方法是表格法和图形法．

1.2.1　频数分布表

通常，统计数据具有一定的规模，如不加以适当分组，则由于过多照顾到细节而不能显示数据分布的趋势和特征；此外，有时研究对象中的各个个体有明显的类型差异，例如，人口调查的数据按年龄段分组可以反映人口的年龄结构、劳动力资源分布和人口老化程度等．因此，对得到的数据进行适当的分组，可以揭示各组的特征和相互关系．

例 1.5　某单位实行弹性上班时间制度，雇员可以在上午 7:00,7:30,8:00,8:30 或 9:00 开始其工作．下面调查了 20 名雇员，得到其开始工作的时间分别为

7:00, 8:30,9:00,8:00,7:30,7:30,8:30,8:30,7:30,7:00,

8:30,8:30,8:00,8:00,7:30,8:30,7:00,9:00,8:30,8:00

由于只有上午 7:00,7:30,8:00,8:30,9:00 这五个开始工作的时刻，因此，可将其数据分成五组．计算每组中数据的频数，得到如下的频数分布表(表 1.3).

表 1.3　　**频数分布表**

开始工作时间	7:00	7:30	8:00	8:30	9:00
频数	3	4	4	7	2

一般地，可将 n 个数据分成 m 组($m\leqslant n$)，第 i 组的频数为 v_i，$i=1,2,\cdots,m$，$\sum_{i=1}^{m} v_i=n$，定义第 i 组的相对频数 f_i 为

$$f_i=\frac{v_i}{n},\quad i=1,2,\cdots,m$$

在上例中可以算出相对频数分布表(表 1.4).

表 1.4　　相对频数分布表

开始工作的时间	7:00	7:30	8:00	8:30	9:00
相对频数	0.15	0.20	0.20	0.35	0.10

一个组的百分数频数＝一个组的相对频数×100.

在上例中,可得其百分数频数分布表(表 1.5)

表 1.5　　百分数频数分布表

开始工作时间	7:00	7:30	8:00	8:30	9:00
百分数频数	15	20	20	35	10

从百分数频数可以清楚看到各组数据按百分数的分布情况. 例如,从上表可以看出:被调查者中选择 8:30 开始工作的比率最高,其次为 8:00 和 7:30.

有些考察对象的指标不是离散指标(例如重量、长度等),此时要先决定分组的组数和每组的组下限、组上限. 规定每组的组中值为

$$\text{组中值}=\frac{1}{2}(\text{组下限}+\text{组上限}).$$

例 1.6　表 1.6[①] 给出 1995 年我国各地区的死亡率(资料来源:《中国人口统计年鉴 1996》).

表 1.6　　我国各地区 1995 年死亡率　　(单位:‰)

地　区	死亡率	地　区	死亡率	地　区	死亡率
北　京	5.12	浙江	6.75	海南	5.61
天　津	6.23	安徽	6.41	四川	7.21
河　北	6.32	福建	5.90	贵州	7.60
山　西	6.12	江西	7.28	云南	8.03
内蒙古	6.70	山东	6.47	西藏	8.80
辽　宁	6.15	河南	6.28	陕西	6.57
吉　林	6.09	湖北	6.91	甘肃	6.49
黑龙江	5.33	湖南	7.15	青海	6.89
上　海	7.05	广东	5.70	宁夏	5.49
江　苏	6.56	广西	6.53	新疆	6.45

① 资料中没有列入我国港、澳、台地区。

将表 1.6 数据按由小到大的顺序排列，发现最低为 5.12‰，最高为 8.80‰. 因此，所有数据均在[4.90,9.10]内，现决定将其分为 6 组（具体分成多少组合适，无具体规定．一般而言，根据数据的个数来定，原则上，数据个数在 50 以下时，分成 5～6 组；数据个数在 50～100 之间时，分成 6～10 组.）. 如果按等距分组，可以算出组距为 $\frac{9.1-4.9}{6}=0.7$，由此可将[4.90,9.10]分成 6 个组：[4.9,5.6)，[5.0,6.3)，[6.3,7.0)，[7.0,7.7)，[7.7,8.4)，[8.4,9.1]. 由表 1.6 给出的数据可得到下列频数和百分数频数分布表（表 1.7）.

表 1.7　　我国各地区 1995 年死亡率频数和百分数频数分布表

分组编号	组　段	频数(v_i)	百分数频数(%)	组中值
1	[4.9,5.6)	3	10.00	5.25
2	[5.6,6.3)	8	26.67	5.95
3	[6.3,7.0)	12	40.00	6.65
4	[7.0,7.7)	5	16.67	7.35
5	[7.7,8.4)	1	3.33	8.05
6	[8.4,9.1]	1	3.33	8.75

有时，根据实际情况，会采用不等距的分组.

例 1.7　在某城镇进行家庭月均收入调查，共调查了 1000 个家庭，根据调查后得到的数据，发现最低为 210 元，最高为 4800 元，建立表 1.8 的频数和百分数频数分布表.

表 1.8　　1994 年某城镇家庭月均收入数据的频数和百分数频数分布表

分组编号	组段(元)	频数(v_i)	百分数频数(%)	组中值(元)
1	200～300	125	12.5	250
2	300～500	170	17.0	400
3	500～800	380	38.0	650
4	800～1000	150	15.0	900
5	1000～1300	86	8.6	1150
6	1300～2000	59	5.9	1650
7	2000～5000	30	3.0	3500

在本例中，如果仍按等距分组，会造成有些组的频数为零，这是应当避免的.

1.2.2 条形图、柱形图和饼形图

条形图和柱形图是两种描述品质数据的频数、相对频数或百分数频数分布的方法.

例 1.5(续) 由例 1.5 中得到的频数分布可建立图 1.1 的**柱形图**.

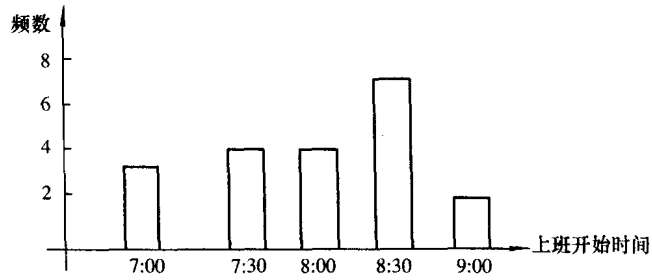

图 1.1 开始工作时间数据的频数的柱形图

柱形图主要用于各项信息的标识(名称)较短的情况.

当各项信息的标识(名称)比较长时,应当尽量使用条形图.

图 1.2 给出了 1996 年日本、韩国、美国和我国港澳台来北京旅游人数的条形图.

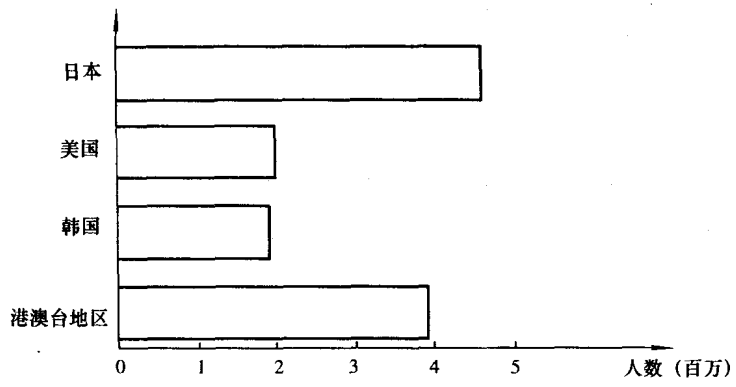

图 1.2 来北京旅游人数的条形图

饼形图一般用来描述和表现各成分占全部的百分比. 例如,各品牌彩电的市场占有率. 我们用圆盘来表示全体,用百分数频数把圆盘剖分成各扇形区域,各扇形区域的面积与其百分数频数相对应.

在例 1.5 中,开始工作的时间分为五种:7:00,7:30,8:00,8:30,9:00,各种情况所占的比例分别为 15%,20%,20%,35%,10%. 用饼形图表示时,各部分扇形的角

度分别为 $0.15\times360^\circ=54^\circ$，$0.20\times360^\circ=72^\circ$，$0.20\times360^\circ=72^\circ$，$0.35\times360^\circ=126^\circ$，$0.10\times360^\circ=36^\circ$，画成饼形图(图 1.3).

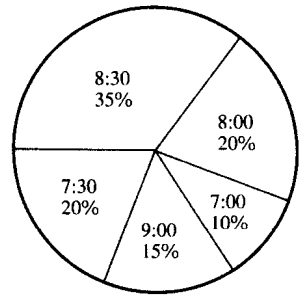

图 1.3　例 1.5 数据的饼形图

1.2.3　散点图

散点图用来描述成对数据中两个变量之间的关系.

表 1.8 给出了阿姆德比萨饼连锁店在各个大学旁开出的连锁店的季度销售收入与学校学生人数的数据.

表 1.8　　比萨饼连锁店的学生人数和季度销售收入数据

连锁店(i)	学生人数(x_i)(千人)	销售收入(y_i)(千美元)
1	2	58
2	6	105
3	8	88
4	8	118
5	12	117
6	16	137
7	20	157
8	20	169
9	22	149
10	26	202

以横坐标表示学生人数，纵坐标表示季度销售收入，将表 1.8 中给出的数据$(x_i, y_i)(i=1,2,\cdots,10)$画在平面上，就得到图 1.4. 图 1.4 称为数据$(x_i, y_i)(i=1,2,\cdots,10)$的散点图. 从图上可以看出：学生人数多的高校的连锁店其销售收入一般也高.

1.2.4　直方图

直方图主要适合于连续变量数据，有了一张频数分布表后，就可着手制作直方图了. 先画出一根水平直线，标明单位长度的实际意义，长短的确定要参考数据的范围，然后在这根水平直线上标出刻度以及频数分布表中的分组段；其次在水平直线上方以每一组段为底边作一矩形，要求每一个矩形的面积等于分布表中该组段数据的百分数频数，因此，该矩形的高 h 为

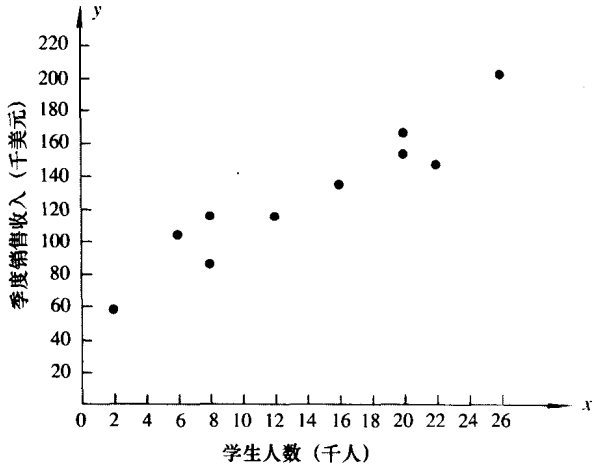

图 1.4　比萨饼连锁店的学生人数和季度销售收入散点图

$$\text{矩形高 } h=\frac{\text{组段内数据的百分数频数}}{\text{组段的长度}}$$

在这里，事先确定一个高度的单位长度更方便. 一般说来，高度单位长度的选择以画出的直方图清楚匀称为好. 通常可事先画一条垂直于水平直线的纵向直线，在其上设定矩形高度的长度单位，然后在纵向直线上标上刻度，使每一单位长度等于水平直线上单位长度的百分数，称这样的纵向刻度为密度尺度. 下面的图 1.5 是例 1.6 数据的直方图.

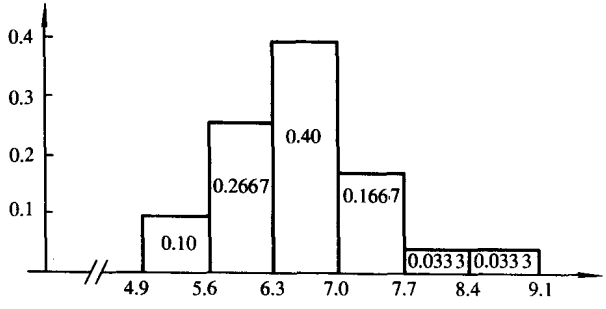

图 1.5　我国 1995 年各地区死亡率的直方图

图 1.5 中各矩形的面积分别为 0.10，0.2667，0.40，0.1667，0.0333 和 0.0333.

例 1.8　20 件机器零件的质量(单位：kg)分别为：215，227，216，192，207，207，214，218，205，200，187，185，202，218，195，215，206，202，208，210. 其最小值为 185，

最大值为227，将其等距地分成5组．表1.9给出了这组数据的频数分布表.

表1.9　　20件机器零件数据的频数分布表

分组编号	1	2	3	4	5
组　段	[184.5,193.5]	(193.5,202.5]	(202.5,211.5]	(211.5,220.5]	(220.5,229.5]
组中值	189	198	207	216	225
百分数频数(%)	15	20	35	25	5

图1.6是本例数据的直方图.

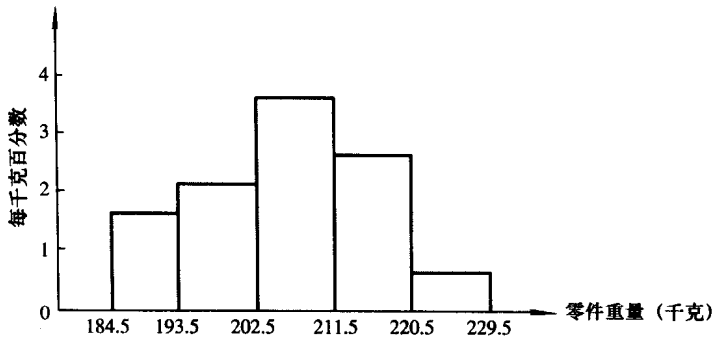

图1.6　20件机器零件质量直方图

用直方图表示的数据分布好像是百分数频数在每个组区间上“均匀散布”，这是分布的一种近似，理论上，这些矩形面积的总和应该等于100%，但实际计算数据的频数分布表中的百分数频数采用四舍五入，因而总和可以与100%略有出入.

对于离散数据，也可用直方图进行描述，但意义不大.

直方图的优点是对数据的分布有一个整体描述，如要求不甚精确，可以在直方图上找出研究对象分布的某些特征．例如，在例1.7中，由直方图可以大致看出：该批零件的质量分布的中心位置大概是在205～206kg，且在其左右出现的可能性大约各占一半，分布呈两头小、中间大的近似正态曲线形态.

第三节　描述统计学Ⅱ：概括统计量

从实际问题中收集到的数据杂乱无章，通过计算得到它的频数分布表，进而画出直方图，这一过程是数据整理的前期工作．直方图可以从整体角度形象地描述收集到的统计数据的大致趋势，是数据分布的一种近似，具有直观和形象的优点．本节我们进一步发掘隐藏在直方图所描述的数据分布现象中的统计规律，即充分利用数据