第 2 章　数据收集与显示

案例导入：

庞大的中国网民队伍

中国互联网络信息中心(CNNIC)2017 年 1 月 22 日发布了第三十九次《中国互联网络发展状况统计报告》。《报告》显示，截至 2016 年 12 月，我国网民规模达 7.31 亿，互联网普及率达到 53.2%，超过全球平均水平 3.1 个百分点，超过亚洲平均水平 7.6 个百分点。

《报告》显示，我国 2016 年全年共计新增网民 4299 万人，增长率为 6.2%，我国网民规模已经相当于欧洲人口总量。其中，手机网民规模达 6.95 亿，占比达 95.1%，增长率连续 3 年超过 10%。而台式计算机、笔记本电脑的使用率均出现下降，手机不断挤占其他个人上网设备的使用。

《报告》显示，2016 年，我国手机网上支付用户规模增长迅速，达到 4.69 亿，年增长率为 31.2%，网民手机网上支付的使用比例由 57.7%提升至 67.5%。手机支付向线下支付领域的快速渗透，极大丰富了支付场景，有 50.3%的网民在线下实体店购物时使用手机支付结算。

《报告》指出，我国网民规模经历近 10 年的快速增长后，红利逐渐消失，网民规模增长率趋于稳定。2016 年，中国互联网行业整体向规范化、价值化发展，同时，移动互联网推动消费模式共享化、设备智能化和场景多元化。

以上资料比较简明扼要地反映了当下中国网民用户的基本情况，问题的关键在于，上述数据是怎么得到的？众多的数据如何显示才会让我们对研究对象的数量特征与规律“一目了然”？

2.1　数据的计量

2.1.1　数据的计量尺度

统计数据是对各种客观现象的信息进行计量的结果，由于现象的性质不同，予以计量的尺度或测量的程度也是不同的。例如，有的现象只能对或只需对其属性进行分类，如人口的性别和民族、产品的质量等级、服务态度的好坏等；有些则可以或要求必须用比较精确的数字加以计量，如一定时期一定地区的经济活动总量、人口的平均寿命、企业的销售收入、学生考试的平均成绩等。根据计量学的一般分类方法，按照对事物计量的精确程度，可以将采用的计量尺度由低级到高级、由粗略到精确分为四个层次，即定类尺度、定序尺度、定距尺度和定比尺度。采用不同计量尺度可以得到不同类型的统计数据，进而需要采用不同的统计分析方法进行分析研究。

1．定类尺度

定类尺度也称类别尺度或列名尺度，是最粗略、计量层次最低的计量尺度。运用“属于

或不属于”的判断标准对事物的某种属性进行平行的分类或分组，用于测量定类变量，如性别分类、民族划分、企业所有制属性等，计量的结果只是表现为某种类型，各个类别之间是平等、并列关系，没有顺序、大小、优劣之分。为了便于统计处理特别是计算机的识别，对不同的类别用自然数字表示或编码表示，例如，用编号“1”表示男，编号“2”表示女；用编号“1”表示汉族，用编号“2”表示苗族，用编号“3”表示满族等，但是这些数字不可以区分大小或进行任何的数学运算，只能计算各个类别或组内的频数或频率。

定类尺度具有的数学特性：= 和 ≠。

2．定序尺度

定序尺度也称顺序尺度，是对事物之间等级差别或顺序差别的一种测度，它不仅能将不同的事物分为不同的类别，还可以确定这些类别的优劣或顺序。一般可以用数字与字符表示，如受教育程度分为大专以上、高中、初中和小学及以下几类，可以分别编号为 1，2，3，4；考试成绩等级可以分为优、良、中、及格、不及格；职称变量可以分为初级、中级、高级，等等。显然，定序尺度对事物的计量比定类尺度要精确一些，这种测量值不仅反映了类别差异还反映了次序差异，但类别之间的准确差值无法说明。因此该计量尺度包括定类尺度(等于或不等于)的数学特性，还具有<和>的数学特性，但还是不能进行加减乘除等数学运算。

定序尺度具有的数学特性：= 和 ≠；>和<。

3．定距尺度

定距尺度也叫间隔尺度，它不仅能把事物分为不同类别并进行排序，还可以准确地计量它们之间的差距。显然，度量的层次高于定序尺度，是一种较精密的计量尺度，通常是使用自然的或物理的单位作为计量的尺度，如长度用米度量、重量用吨度量、考试成绩用百分制度量、收入用人民币元度量等。定距尺度的计量结果表现为具体的数值。由于这种尺度的每一间隔都是相等的，因此只要给出一个度量单位，就可以准确地标出两个数值之间的差值。比如，甲乙两地的海拔高度分别是 2000 米和 1000 米，甲乙两地的温度分别是 30 摄氏度和 20 摄氏度，甲乙学生的考试成绩分别是 85 分和 70 分，比较它们的顺序和异同，并计算其差距，即甲地的海拔高于乙地 1000 米，甲地的温度高于乙地 10 摄氏度，学生甲的分数高于学生乙 15 分。 因此定距数据不仅具有定类尺度和定序尺度的特性，其结果还可以进行加减运算。

定距尺度具有的数学特性：= 和 ≠；>和<；+ 和 −。

4．定比尺度

定比尺度与定距尺度属于同一层次，也是用数值对现象进行计量的较为精密的计量尺度，也称比率尺度。定比尺度除了具有上述三种计量尺度的全部特性以外，还具有一个特性，就是可以计算两个测度值之间的比值，这就要求定比尺度中有绝对“零值”，所谓“零值”或“0”表示“没有”或“不存在”，这是它和定距尺度的唯一差别。

定比尺度具有的数学特性：= 和 ≠；>和<；+ 和 −；×和÷。

在定距尺度中没有绝对零点，如果定距尺度的计量值为 0，表示一个数值，即“0”水平，而不表示“没有”或“不存在”，如成绩 0 分、海拔 0 米、温度 0 摄氏度，不能说没有成绩、没有海拔高度、没有温度。而在定比尺度中，“0”表示“没有”或“不存在”，这犹如现实生活中的“0”大多表示“无”、“没有”或“不存在”的意思，如职工人数、销售收入、固定资

产投资额、移动电话户数等为 0，表示该事物不存在。因此对于定比变量，除了可以分类、比较大小、进行加减乘除以外，还可以计量测度值之间的比值，职工人数 200 人比职工人数 100 人多 100 人，还可以计算人数 200 人是人数 100 人的 2 倍，但考试成绩 90 分不能计算说是分数 45 分的 2 倍。

上述 4 种计量尺度对事物的计量层次是从低级到高级、从粗略到精确逐步递进的。高层次的计量尺度包括低层次的计量尺度的所有特性，如表 2.1 所示；高层次计量尺度的计量结果可以很容易地转化为低层次计量尺度的计量结果，反之则不那么容易。如将不同数值的业务收入容易转化为高、中、低业务收入，而将高、中、低业务收入转化为具体数值的业务收入几乎不可能。

表 2.1 四种计量尺度的比较

	定类尺度	定序尺度	定距尺度	定比尺度
分类(=，≠)	√	√	√	√
排序(>，<)		√	√	√
差值(+，–)			√	√
比值(×，÷)				√

由于高层次的计量尺度包括低层次的计量尺度的所有特性，对事物的计量更精确，可以应用的统计分析方法更多，分析也更方便，因此，在统计分析中，应尽可能使用高层次的计量尺度。

2.1.2 数据的类型

与数据计量尺度相对应，数据也有 4 种：定类数据、定序数据、定距数据、定比数据。将定类数据和定序数据称为品质数据或定性数据，将定距数据和定比数据称为数值型数据或定量数据。

统计数据是利用某种计量尺度对现象进行计量的结果，采用不同的计量尺度将得到不同类型的数据。计量尺度有上述 4 种，相应地，统计数据也有 4 种类型。定类数据，由定类尺度计量形成；定序数据，由定序尺度计量形成；定距数据，由定距尺度计量形成；定比数据，由定比尺度计量形成。

定类数据和定序数据均表现为类别，是说明事物的品质、属性特征的，称为品质数据或定性数据。品质数据一般用文字表达，不用数值表示，即使用数值表示，也只是表示其符号或代码，数值本身不表明计算结果的大小。如对经济活动的不同区域计量形成的数据为“东部”、“中部”或“西部”，便于处理可分别用编号“1”、“2”、“3”表示，但并不表示“西部”要比“东部”大。

定距数据和定比数据均表现为数值，是说明事物数量特征的，称为数值型数据或定量数据、数量数据。定量数据的数值表示计量结果的大小。不同类型的数据，必须采用不同的统计方法进行分析与处理。统计学中研究的统计数据主要是数量数据。

区分计量的层次和数据的类型是十分重要的，因为对不同类型数据将采用不同的统计方法来处理和分析，见表 2.2。

表 2.2　不同计量层次、不同数据类型与不同统计分析方法比较

测量尺度	数据类型	一般案例	适用的统计分析方法	
			描述统计方法	推断统计方法
定类尺度	类型数据	单位性质	比例、众数、异众比率	列联表分析、卡方检验等
定序尺度	顺序数据	质量等级	比例、中位数、四分位差	计算等级相关系数等非参数分析
定距尺度	数值型数据	温度	全距、均值、标准差	积差相关系数、t 检验、ANOVA 回归、因子分析
定比尺度	数值型数据	重量	几何均值、调和平均数	变异系数

2.2　数据的收集

数据的收集就是进行统计调查。统计调查是根据统计研究的目的和要求，有组织、有计划地向调查对象收集原始资料和次级资料的过程。原始资料也称第一手资料或初级资料，调查者通过实地调查或亲自试验所获得；次级资料也称二手资料，是调查者借用别人已经加工整理过的资料，如从统计年鉴、会计报表或其他的出版刊物上所获得的资料。统计调查所收集的资料主要指原始资料。

2.2.1　数据的直接获取

1．普查

普查是为某一特定目的而专门组织的一次性的全面调查。它对调查总体中的每一个个体单位都要进行调查，如人口普查、经济普查、农业普查、疾病普查等。

普查是适合于特定目的、特定对象的一种调查方式，主要用于收集处于某一时点状况上的社会经济现象的数量，目的是掌握特定社会经济现象的基本全貌，摸清和掌握有关国情国力、企业基本实力等基本情况，为国家、部门或企业制定有关政策、措施或决策提供依据。

普查具有以下特点：

(1) 普查有一次性的也有周期性的。普查按一定周期进行，便于研究现象的发展趋势及其规律性。如人口普查每 10 年进行一次，尾数逢 0 的年份为普查年度。新中国成立以来，我国已经成功进行过五次全国人口普查，分别在 1953 年、1964 年、1982 年、1990 年和 2000 年，2010 年 11 月 1 日开始了第六次人口普查。在 2004 年开展第一次全国经济普查，每 10 年进行两次，分别在逢 3、逢 8 的年份实施。农业普查每 10 年进行一次，每逢 7 的年份进行。

(2) 规定普查的项目和指标。普查时必须按照统一规定的项目和指标进行登记，不准任意改变或增减，以免影响汇总和综合，降低资料质量。同一种普查，每次调查的项目和指标应力求一致，以便于进行历次调查资料的对比分析和观察社会经济现象发展变化情况。

(3) 普查的规范化程度高，所获得的资料全面，其数据一般比较准确。因此它可以为抽样调查或其他调查提供基本依据。如人口普查登记的主要内容：姓名、性别、年龄、民族、户口登记状况、受教育程度、行业、职业、迁移流动、社会保障、婚姻、生育、死亡、住房情况等。

(4) 普查使用范围比较窄，只能调查一些最基本的、特定的现象。

(5) 普查一般需要规定统一的标准调查时间。标准调查时间包括：资料所属的时间即标准时点，调查工作期限。

标准时点是指对被调查对象登记时所依据的统一时点。调查资料必须反映调查对象的这一时点上的状况，以避免调查时因情况变动而产生重复登记或遗漏现象。例如，我国人口普查的标准时点为普查年份的 11 月 1 日零时，就是要反映这一时点上我国人口的实际状况；农业普查的标准时点定为普查年份的 1 月 1 日 0 时；经济普查的标准时点为普查年份的 12 月 31 日。

而调查工作期限是指，在普查范围内各调查单位或调查点尽可能同时进行登记，并在最短的期限内完成，以便在方法和步调上保持一致，保证资料的准确性和时效性。如第六次人口普查 2010 年 11 月 1 日—10 日为入户登记；11 月 16 日—12 月 25 日为事后质量抽查；2011 年 3 月至 2012 年 5 月期间公布普查数据；2010 年 12 月至 2012 年 6 月进行普查资料开发利用和分析。

(6)普查工作量大，花费大，组织工作复杂。

2．抽样调查

抽样调查是根据随机原则从调查总体中抽取部分单位进行观察并根据其结果推断总体数量特征的一种非全面调查方法。

抽样调查是实际中应用最广泛的一种调查方式和方法，具有以下几个特点：

(1)经济性。由于调查的样本通常是总体单位中很少的一部分，调查的工作量小，因而可以节省大量的人力、物力、财力和时间。

(2)时效性。由于整体工作量小，调查准备时间、调查时间、数据处理时间等环节都可以大大缩减，从而提高数据的时效性，可以迅速及时地获得所需要的资料。

(3)准确性。因为全面调查的工作量大，环节多，登记性误差往往很大；而抽样调查由于工作量小，可以把各个环节的工作做得更细致，登记性误差往往很小。因此抽样调查的数据质量有时比全面调查更高。

(4)灵活性。由于工作量小、成本低、组织方便，因此可以根据需要对时间、调查对象和调查内容进行灵活的调整。

(5)适应性。由于具备以上 4 个方面的特点，抽样调查广泛应用于社会经济和科学技术的各个领域，能够解决全面调查无法或很难解决的问题。对于无法进行全面调查或不可能进行全面调查以及没有必要进行全面调查的总体，抽样调查就可以发挥其作用。

2.2.2 数据的间接获取

数据的间接来源是指次级资料(二手资料)的收集，次级资料主要是公开出版或公开报道的数据，当然有些是尚未公开的数据。次级资料的使用，有些是免费的，有些是有偿的。如果统计资料通过某些渠道可以来源于已有的数据，就无须花费大量的人力、时间和费用进行直接调查。

1．公开出版和公开报道的数据

(1)政府出版物

政府出版物是数据来源的主要渠道。中国政府提供的统计数据资料范围最广、信息最多。如由国家统计局编辑、国家统计出版社出版发行的《中国统计年鉴》、《国民经济和社会发展统计公报》、各部、委公开出版物等。在《中国统计年鉴》中，有每年中国的国民经济和社会

发展情况的各方面的数据资料，包括人口和劳动力、农业、工业、运输、邮电、固定资产投资、商业、对外贸易和旅游、财政金融、物价、人民生活、教育科学文化、科技发明创造、体育卫生等资料。

(2)法人组织出版物

法人组织出版物的统计资料或调查数据资料一般通过出版社来公开出版发行，如各种经济信息中心、信息咨询机构、专业调查机构等编制和出版的统计资料，贸易和行业组织公开提供的二手资料，以及来自各类专业期刊、报纸和书籍的资料，等等。

(3)国际组织出版物

联合国、世界银行等每年出版的《统计年鉴》中，有 160 个以上的国家或地区的统计数据。除了有年度出版物，还有月度出版物，如联合国的《统计月报》。其他国际出版物还有《世界经济年鉴》《国外经济统计资料》《世界发展报告》等。

2．非公开出版和报道的数据

这类数据是指在本系统、本行业、本单位内部，用于指导工作、交流信息的非卖性内部资料，不包括机关公文性的简报等信息资料。被纳入公司的内部数据库的各种文件档案：年度报表、股东报告、可向新闻媒介透露的产品测试结果，以及由公司内部有关部门编写的与员工、顾客和其他人员交流的公司刊物等。

3．互联网公布的数据

当今互联网络已成为人们获取信息的非常重要的渠道，几乎所有的政府机构和大公司都有自己的网站并提供公共访问端口，访问者可以从中获得有用的数据。若干统计数据来源网站如表 2.3 所示。

表 2.3　统计数据来源部分网站

各政府机构、公司组织	网　　址	数 据 内 容
中华人民共和国国家统计局	http://www.stats.gov.cn	统计年鉴、统计月报等
国务院发展研究中心信息网	http://www.drcnet.com.cn	宏观经济、财经、货币金融等
中国经济信息网	http://www.cei.gov.cn	经济信息及各类网站
华通数据中心	http://data.acmr.com.cn	国家统计局授权的数据中心
中国互联网络信息中心	http://www.cnnic.net.cn/	互联网发展研究、互联网数据
美国预算编制办公室	http://www.whitehouse.gov/com	财政收入、支出、债券等
美国市场学会	http://www.ama.org	使用关键词可以查询该组织的所有出版物
世界之见	http://www.worldopinion.com	拥有数千份的市场调研报告

相对资料的直接获取，二手资料的收集比较容易，采集数据的成本低，能很快得到。二手资料的作用很大，一般我们在分析研究问题的时候，首先都是从对二手资料的分析开始的。但是二手资料也有其局限性，任何二手资料都是因为其使用者特定的研究问题和研究目的而存在的，在使用二手资料的时候，要注意数据的定义、含义、计算口径和计算方法的变化，避免错用、误用和滥用。并且在引用二手资料的时候要注明数据出处，尊重他人的劳动成果。

2.2.3　数据的质量

统计数据质量的好坏，决定着管理决策的科学性与可靠性。统计的整个工作过程就是对

数据的加工过程，从原始数据的收集开始，经过整理、显示、样本信息的提取到总体数量特征的推断，都涉及统计数据的质量控制问题。统计调查阶段就是收集统计数据，是统计研究的第一步，因此该阶段的数据质量直接影响到整个统计工作过程的质量。

1．统计数据的误差

统计数据的误差，就是调查所得的统计数字与调查总体实际数量之间的离差。例如，对某地区的工业增加值进行调查的结果为 30 亿元，而该地区的工业增加值实际为 29 亿元，那么， 统计调查误差就是 1 亿元。统计数据的误差主要有以下几种：

(1) 登记性误差

登记性误差是由于错误登记事实而发生的误差，不管是全面调查或是非全面调查都会产生登记性误差。分类：偶然性登记误差和系统误差。偶然性登记误差是因为调查人员责任心不强、业务技术水平等原因造成的观察、测量、计算错误、笔误、错填、遗漏，以及被调查者的回答错误所引起的误差。该误差在数量上不会偏向某一方，不具有倾向性。而系统误差是调查人员或被调查者故意虚报、瞒报、假报、故意歪曲事实所引起的误差，具有明显的倾向性，在数量上往往偏向某一方。如“富瞒穷虚”的统计现象，不少富裕地区没有完善全面反映社会经济的发展情况，反映总量指标时瞒的成分非常大，美言“留有余地”，而且在反映增长速度时大搞“橡皮筋”游戏，有很好的伸缩力，想减缓速度少报几个单位，想加快速度多挖潜几个单位；而穷的、经济基础比较薄弱的地区，有强烈的“赶超”意识，千方百计地利用统计上的“盲点”，提高经济总量及其发展速度。

(2) 代表性误差

代表性误差只有非全面调查中才有，全面调查不存在这类误差。非全面调查由于只对调查现象总体的一部分单位进行观察，并用这部分单位算出的指标来估计总体的指标，而这部分单位不能完全反映总体的性质，它与总体的实际指标会有一定差别，这就产生了误差。

当采用抽样调查时，应严格遵守随机原则，保证足够的样本容量，选择适当的抽样调查方式方法，以控制误差的范围。

2．统计数据的质量要求

传统的统计数据质量仅仅指其准确性，通常用统计估计中的误差来衡量。随着人们质量观念的变化，质量的含义不断延伸，准确性已不再是衡量统计数据质量的唯一标准。从用户使用的角度来看，对数据的质量要求主要有以下几个方面。

(1) 准确性：最小的非抽样误差或偏差。

(2) 精度：最低的抽样误差或随机误差。

(3) 时效性：在最短的时间内取得并发布数据。

(4) 适用性：满足用户决策、管理和研究的需要。

(5) 可解释性：发布数据的支持数据以及必要的文字介绍与诠释。

(6) 可比性：加工整理后的数据间要满足可比性。

(7) 完整性：调查的单位无遗漏、调查的内容要齐全。

(8) 最低成本：在满足以上标准的前提下，考虑减轻调查负担，以最经济的方式取得数据。

2.3　数据的显示

2.3.1　数据的审核

对统计调查所获得的数据，在进行整理显示之前，还必须进行严格的审核，以确保统计数据的质量要求。

对数据的审核主要从资料的准确性、及时性和完整性等几个方面进行。

1．直接来源数据的审核

完整性：是否有遗漏，是否填写齐全。

准确性：数据是否真实反映客观实际情况；数据是否有错误，计算是否正确。

2．间接来源数据的审核

对间接来源数据审核的主要内容包括资料的完整性、准确性、适用性和时效性。对统计资料准确性的审核是统计审核的重点，有逻辑检查和技术性检查。逻辑检查是用来检查调查表或报表中的内容是否合理，有关项目之间是否矛盾的一种方法。这种方法要求检查人员坚持实事求是的科学态度，熟悉业务，有一定的实际工作经验和周密的逻辑推理能力。技术性检查主要包括：填报的单位有无遗漏与重复；调查的内容填写是否齐全，所填内容与表格规定是否一致，有无错行与错栏的情况；计量单位是否和法定计量单位一致；各行与各栏间数字的合计项、乘积项等与分项数字是否符合等。

2.3.2　统计分组与频数分布

统计整理的中心任务是统计分组和编制频数分布表。

1．统计分组

(1)统计分组的概念

统计分组就是根据统计研究的需要，将统计总体按照一定的标志区分为若干组成部分的一种方法。通过分组把现象内部不同性质或不同数量的单位分开，按性质或数量相同的单位归并在一个组内，说明现象内部各组之间的相互关系及其特征。例如，大学生按性别、年龄和政治面貌分组，工业企业按经济类型、生产规模大小和所属行业分组，等等。

(2)统计分组的标志

进行统计分组时，最关键的问题是如何选择分组的标志和确定各组的界限。所谓分组标志，就是将总体区分为不同组别的标准或依据。分组标志有数量标志和品质标志两种。一个总体一般具有多种特征，对同一资料采用的分组标志不同，有可能得出相异甚至相反的结论。

分组的基本原则是按照不同的标志分组，体现组内的同质性和组间的差异性。

① 按品质标志分组。就是按事物的品质特征进行分组。如企业按经济类型分组，可以有国有企业、集体企业和其他类型；按管理系统分组，可以分为中央直属企业、地方所属企业。人口按性别可分为男和女两组；按民族可分为汉族、蒙古族、朝鲜族、回族、藏族和壮族等。

② 按数量标志分组。就是按事物的数量特征进行分组。例如，居民家庭按子女数分组，

企业按工人数、产值和资产数量等标志进行分组。按数量标志分组，不仅可以反映事物数量上的差别，有时通过事物的数量差异也可以区分事物的性质。如学生按考试成绩分组：60 分以下，60～70 分，70～80 分，80～90 分，90～100 分。企业按计划完成程度分组：100%以下，100%～110%，110%～120%，120%～130%，130%以上。前一种分组可以分析学生的学习成绩是否及格以及成绩水平；后一种分组可用于分析企业是否完成计划以及完成的好坏。

2．频数分布

(1)频数分布的概念

在统计分组的基础之上，将总体中的所有单位按组归类整理，并按一定的顺序排列，形成总体单位在各组间的分布，称为次数分布、频数分布或分布数列。分配在各组的总体单位数叫作次数，又称频数；各组次数占总次数之比称为频率。

根据分组标志的不同，分布数列可分为属性分布数列和变量分布数列两种。

属性分布数列是指按属性标志分组所形成的分布数列。例如，工业企业按行业、所有制形式、属地分组，人口按性别、籍贯、受教育程度等分组。对于属性分布数列来讲，如果分组标志选择得好、分组标准定得恰当，则事物的差异能明确地被表现出来，各组的划分就很容易，并且分组的结果所形成的属性分布数列一般较稳定，总体的分布特征能够准确反映出来。我国大陆 2016 年年末人口分布数列如表 2.4 所示。

表 2.4　我国大陆 2016 年年末人口性别分布的属性分布数列

按性别分组	人数(万人)	比率(%)
男	70 815	51.2
女	67 456	48.8
合计	138 271	100.00

变量分布数列是指按数量标志分组形成的分布数列。分组的结果因人而异，按同一数量标志分组有可能出现多种分布数列结果。

变量分布数列按照分组变量的表现形式，可以分为单项式变量数列和组距式变量数列。

① 单项式变量数列。单项式变量数列中每个组的变量都只有一个，即一个变量值代表一组，适合于变异幅度不大的离散型变量分组。大二某班 30 名学生年龄分布数列如表 2.5 所示。

表 2.5　某班同学年龄分布的单项式变量数列

按年龄分组(岁)	人数(人)	比率(%)
18	3	10.00
19	8	26.67
20	15	50.00
21	3	10.00
22	1	3.33
合计	30	100.00

② 组距式变量数列。组距式变量数列是按一定的数量变化范围或距离分组的结果，又称组距数列。每组中距离的大小称为组距，即：组距=上限–下限。根据每组组距是否相同，组距式数列又有等距数列与不等距(异距)数列之分。组距式数列适合于变量个数较多，数量变化范围大的资料。某班学生统计学期末考试成绩分布等距数列如表 2.6 所示。

表 2.6　某班统计学期末考试成绩分布的等距数列

按成绩分组	人数(人)	比率(%)
50～60	4	8
60～70	9	18
70～80	19	38
80～90	11	22
90～100	7	14
合计	50	100

注：上表分组中的上、下限重叠时，一般原则是达到上限值的单位划入下一组内，即“上限不在内”，如 70 分计入第 3 组。

2015 年《中国统计年鉴》显示了我国 2014 年的人口年龄构成，如表 2.7 所示的异距数列。

表 2.7　人口按年龄阶段分组的异距数列

年龄(岁)	人数(万人)	百分比(%)
0～14	22 558	16.50
15～64	100 469	73.40
65 以上	13 755	10.10
合计	136 782	100.00

资料来源：中华人民共和国国家统计局

(2) 频数分布表的编制

编制频数分布表的步骤是：

第一，整理原始资料，计算全距。

第二，确定变量数列形式。

第三，组距式变量数列的编制。

编制组距式变量数列主要需解决四个问题：一是组距、组数的确定；二是组距数列的形式；三是组限的确定；四是各组次数的计算。

① 组距、组数的确定。组数=全距÷组距。组距大小要合适，要能正确反映总体的分布特征及其规律。组距过大，组数过少，容易把不同质的单位归在一个组内；组距过小，组数过多，又容易把同质的单位分在不同组内。两者都不符合分组的要求。如果总体呈正态或近似正态分布的情况下，组距可以根据皮尔逊经验公式给出：

$$\text{组距} = \text{全距}/(1+3.322\log_{10}n)$$，n 为观察值个数

② 等距数列、异距数列的选择。至于采用等距还是异距分组，要根据现象的特点、统计研究的目的以及收集到的资料分布的均匀性来确定。许多社会经济现象的数量变化呈现正态或近似正态分布的情况，适于采用等距分组形成等距数列。有时同一现象，如在研究人口年龄构成时，可以根据研究目的的不同，采用等距式或异距式来划分年龄组。通常划分年龄层次的方法有四种：

基本年龄组。又叫一岁年龄等距分组，即以 1 岁为组距，不满周岁为 0 岁组，满 1 周岁不到 2 周岁为 1 岁组，……，依次类推。

常见年龄组。按 5 岁或 10 岁为组距将人口进行等距分组。如按 5 岁分，则可以划分成 0～4 岁组，5～9 岁组，10～14 岁组，……，依次类推。

主要年龄组。采用国际通用的三种主要年龄为界限将人口划分为幼年组(0～14 岁)，成年组(15～64 岁)，老年组(65 岁以上)。

特殊年龄组。从需要出发，根据人口的各种社会经济特征把人口划分为若干年龄层次。如 0 岁为婴儿组；1～6 岁为学龄前儿童组；男 18～60 岁、女 18～55 岁为劳动年龄组；15～49 岁(女性)为有生育能力人口组，等等。

我国人口按年龄阶段的划分，如表 2.7 所示，是根据联合国提出的标准进行的不等距分组；如果观察年龄性别构成或编制生命表，可以用 5～10 岁作组距进行等距分组，如反映人口数量特征和变化趋势的人口年龄金字塔就是一个等距分组数列，如表 2.8 所示为人口年龄变化趋势的等距数列。

表 2.8　人口年龄变化趋势的等距数列

年龄(岁)	人口数(人)		人口数占总人口比重(%)		性　别　比
	男	女	男	女	(女=100)
合计	576 011	548 391	51.23	48.77	105.04
0～4	34 484	29 506	3.07	2.62	116.87
5～9	34 326	28 807	3.05	2.56	119.16
10～14	31 616	26 671	2.81	2.37	118.54
15～19	34 584	30 136	3.08	2.68	114.76
20～24	46 891	43 894	4.17	3.90	106.83
25～29	49 801	49 044	4.43	4.36	101.54
30～34	41 777	40 768	3.72	3.63	102.47
35～39	41 761	40 032	3.71	3.56	104.32
40～44	52 086	49 873	4.63	4.44	104.44
45～49	50 455	48 795	4.49	4.34	103.40
50～54	39 470	38 439	3.51	3.42	102.68
55～59	33 781	32 628	3.00	2.90	103.54
60～64	30 781	30 826	2.74	2.74	99.85
65～69	20 573	21 137	1.83	1.88	97.33
70～74	14 528	14 606	1.29	1.30	99.47
75～79	10 179	11 151	0.91	0.99	91.28
80～84	5987	7302	0.53	0.65	81.99
85～89	2244	3360	0.20	0.30	66.79
90～94	581	1175	0.05	0.10	49.45
95 以上	106	241	0.01	0.02	43.98

表 2.8 依据 2014 年全国人口变动情况抽样调查(抽样比为 0.822‰)数据整理。资料来源于《中国统计年鉴》(2015)。

③ 组限的确定。上限与下限统称为组限。确定组限的基本原则是：按这样的组限分组后，要能使性质相同的单位归入同一组内，性质不同的单位按不同的组别划分。

对于离散型变量，其变量值都是整数，组的上下限可用肯定性的数值表示，相邻组的上下限一般不重叠。

例如，企业按职工人数分组，其组限可表示为：100 人以下，100～299 人，300～499 人，500～699 人，700 人以上。

对于连续型变量，其变量值有小数，组限不能用肯定的数值表示，所以相邻组的组限必须重叠。

例如，企业按单位职工工资分组，组限可以表示为：100 元以下，100～300 元，300～500 元，500～700 元，700 元以上。

上述组限的表达中，上下限齐全的叫闭口组，缺一个组限的叫开口组。

④ 各组次数的计算。下面举例说明各组次数的计算方法。

【例 2.1】 以下是一个班级 50 名学生统计学考试成绩资料，编制组距式变量数列。

30　79　87　88　89　65　62　60　63　78　78　84　67　68　69　67　89　90　79
98　95　76　56　91　90　86　81　78　79　76　67　78　79　70　45　56　78　79
98　97　87　86　84　79　76　75　73　72　86　75

解：首先，将这些资料按一定顺序进行排列并确定全距。通过使用 Excel 排序，计算全距=98–30=68。最低分数与最高分数相差 68 分，有一个极端分数出现，除去 4 个分数小于 60 分以外，其余都分布在 60～100 之间。

其次，确定数列的类型。根据上一步分析，编制组距式数列。

最后，确定组距和组数。组距定为 10 分，组数定为 5 组，各组依次表现为 60 分以下，60～70 分，70～80 分，80～90 分，90～100 分，形成分布数列，如表 2.9 所示。

表 2.9　学生成绩次数及累积累计次数分布表

成绩(分)	人数(人)	频率(%)	向上累计频数(人)	向下累计频数(人)
60 以下	4	8	4	50
60～70	9	18	13	46
70～80	19	38	32	37
80～90	11	22	43	18
90～100	7	14	50	7
合计	50	100	—	—

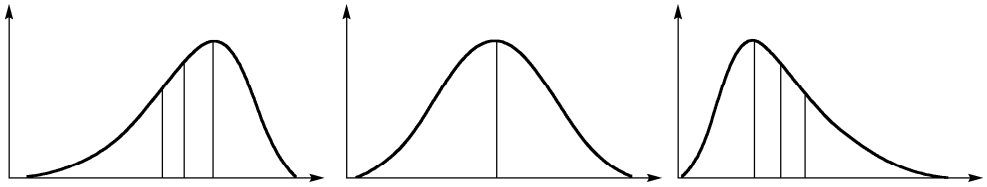
为了统计分析的需要，有时要观察某一数值变量以上或以下的次数之和，这时要计算累计次数，编制累计次数分布表，如表 2.9 所示。根据累计的方向，有向上累计和向下累计。“向上累计”就是从变量值低的向变量值高的方向把分布的次数依次累计相加，反之，则是“向下累计”。如表中的向上累计结果“43”表示 90 分以下的人数有 43 人，而该组的向下累计结果“18”则表示 80 分以上的人数有 18 人。

(3) 频数分布的类型

频数分布的形式常表现为钟形分布、U 形分布和 J 形分布。

① 钟形分布。钟形分布的特征是“两头小，中间大”，即靠近中间的变量值分布的次数多，靠近两边的变量值分布的次数较少，其曲线图宛如一口钟，如图 2.1(b)所示。其分布特征是以变量的均值为对称轴，左右两侧对称，两侧变量值分布的次数随着与均值的距离的增大而逐次减少。在统计学中称为正态分布，是对称钟形分布。在自然和社会现象中，大量随机变量都服从或近似服从这种分布，如人类的身高、测量某零件长度的误差、学生成绩，等等。而图 2.1(a)(c)属于偏态分布，分左偏和右偏，居民收入一般呈现右偏分布，而一个老龄化社会的人口年龄分布则呈现左偏分布。

② U 形分布。U 形分布与钟形分布的图形相反，即“两头大，中间小”，靠近中间的变量值次数较少，而靠近两边的变量值次数较多。人和动物的死亡率分布大多服从或近似服从 U 形分布，如图 2.2 所示。



(a)左偏分布　　(b)正态分布　　(c)右偏分布

图 2.1　钟形分布

③ J 形分布。J 形分布包括正 J 形分布和反 J 形分布。正 J 形分布表示次数随着变量值增加而增加的现象，如经济学中的供给曲线；反 J 形分布表示次数随着变量值减少而减少的现象，如经济学中的需求曲线。J 形分布如图 2.3 所示。

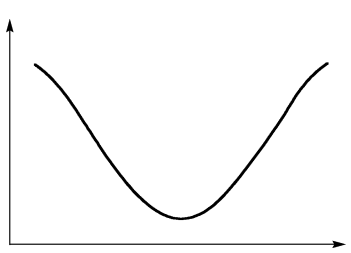
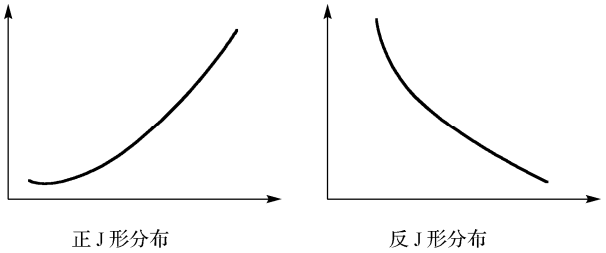


图 2.2　U 形分布



正 J 形分布　　反 J 形分布

图 2.3　J 形分布

2.3.3　数据的显示

1. 统计表

所谓统计表是把统计数据按照一定的结构和顺序用表格显示出来的一种形式。它既是调查整理的工具，又是分析研究的工具。广义的统计表包括统计工作各个阶段中所使用的一切表格，如调查表、整理表和计算分析表等，它们是用来提供统计资料的重要工具。

(1)统计表的构成

从内容结构看，统计表由主词和宾词两个部分组成；从形式结构看，统计表包括总标题、分标题和表中的数字。

从内容上看，统计表由主词和宾词两个部分组成。主词是统计表所要说明的总体及其分组，通常排列在表的左方，列于横栏；宾词是说明总体的统计指标，排列在表的右方，列于纵栏。当然，有时为了更好地编排表的内容，主词和宾词的位置也可以调换。

从构成要素看，统计表包括以下三部分：

① 总标题，统计表的名称，简要说明全表的内容，一般列于表的上端中央。

② 分标题(也称标目)，总体名称或分类名称及说明总体的各种项目。横行标题(也称横栏目)写在表的左方，纵栏标题(也称纵栏目)写在表的上方。

③ 纵横栏组成的本身及表中的数字。

另外，还应有必要的附注和注明资料来源。

下面以表 2.10 为例说明统计表的构成。

由表 2.10 可知，总标题是“2014 年全国分行业增加值”；横行标题是对总体进行的分组，即主词；其他各栏是反映总体规模和说明总体数量特征的统计指标，即宾词。

表2.10　2014年全国分行业增加值　总标题

项　目	增　加　值	
	产值（亿元）	比重（%）
第一产业	58 336.1	9.2
第二产业	271 764.5	42.7
第三产业	306 038.2	48.1
合计	636 138.7	100

横行标题（横标题）　纵栏标题（纵标题）　数字资料

资料来源：《中国统计年鉴》(2015)。

(2)统计表的编制原则

编制统计表总的要求包括：简练、明确、实用、美观、便于比较。在设计时应该注意以下几个方面：

① 表的上下两端用粗线，左右两边不封口；纵栏之间用细线分开，横行之间可以不加线。如果横行过多，也可以每五行加一细线。

② 合计栏的设置。统计表格纵列若需合计时，一般应将合计列在最后一行。各横行若需合计时，可将合计列在最前一栏或最后一栏。

③ 统计表的各种标题，特别是总标题的表达，应十分简明扼要、确切，概括地反映出表的基本内容。总标题还应标明资料所属的时间和地点。

④ 如果统计表的栏数较多，通常要予以编号。主词栏采用甲、乙、丙等文字编号，宾词栏常用数字编号。

⑤ 表中的数字应该填写整齐，对准位数。当数字为 0 或因数小可略而不计时，要写上 0；不应填写的数字的空格用“—”表示；未发生的数字空着不填；估算的数字应在表下说明；无法取得的资料用“…”表示；如果某项数字与邻项数字相同，仍应填写数字，不得用“同上”、“同左”等字样代替。

⑥ 统计表中必须注明数字资料的计量单位。当全表只有一种计量单位时，可以把它写在表头的右上方。如果表中需要分别注明不同单位，横行的计量单位可以专设一栏；纵栏的计量单位要与纵栏标目写在一起，用小字标明。

⑦ 统计表的资料来源及其他需要说明的问题，可在表下加以注明。

2．统计图

前面讲述了用统计表来反映数据资料的频数分布，如果用图形来显示频数分布，结果会更加形象与直观。统计图的类型很多，有平面的，也有立体的，图形均可以由计算机来完成。下面根据数据类型的不同分别介绍各类图形。

定性数据和定量数据采用的统计图通常有所不同。对定性数据常用条形图、饼图、环形图、累计频数分布图等；对定量数据常采用直方图、折线图、曲线图、茎叶图、箱线图、累计频数分布图等。

(1)定性数据的图形显示

① 条形图。条形图是使用宽度相同的条形的高度或长短来表示数据变动的图形。可以横置或纵置，纵置时也称为柱状图。条形图有单式、复式和叠加等形式。

单式条形图。2015 年我国企业营销推广渠道使用情况条形图如图 2.4 所示。

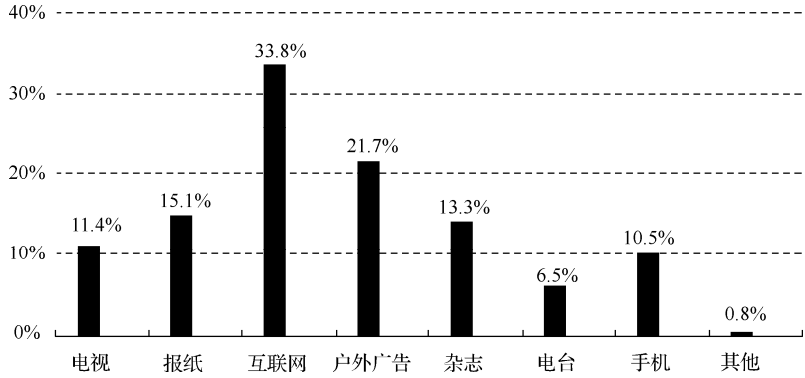


图 2.4 我国企业营销推广渠道

资料来源：http://www.cnnic.net.cn/

2015 年我国企业互联网营销渠道使用比例条形图如图 2.5 所示。

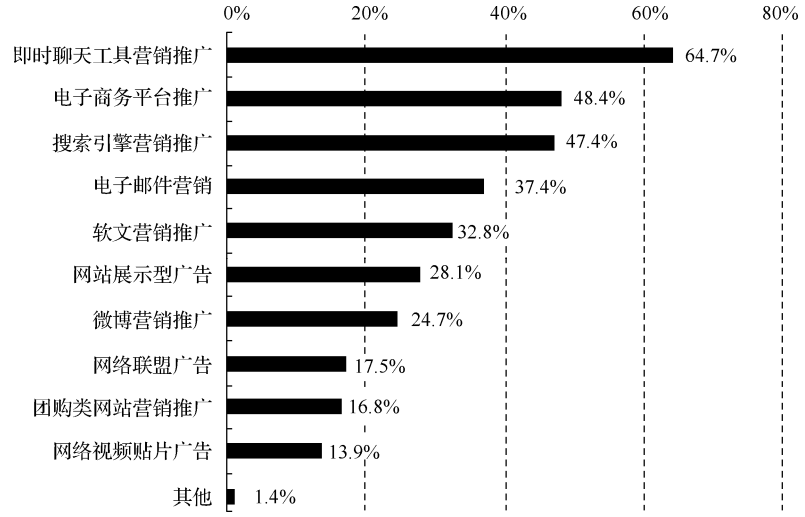


图 2.5 我国企业互联网营销渠道

资料来源：http://www.cnnic.net.cn/

复式条形图，用于多组数据的比较。根据表 2.11 所示资料用 Excel 作复式条形图，如图 2.6 所示。

表 2.11 三大产业就业人员构成

年 份 / 产 业	1990 年	2000 年	2010 年	2014 年
第一产业	60.10%	50.00%	36.70%	29.50%
第二产业	21.40%	22.50%	28.70%	29.90%
第三产业	18.50%	27.50%	34.60%	40.60%

资料来源：《中国统计年鉴》(2015)。

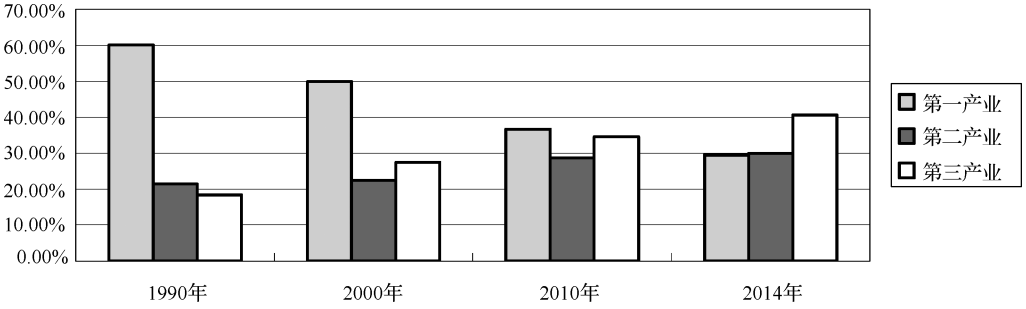


图 2.6　三大产业就业人员构成复式条形图

叠加式条形图，如图 2.7 所示。

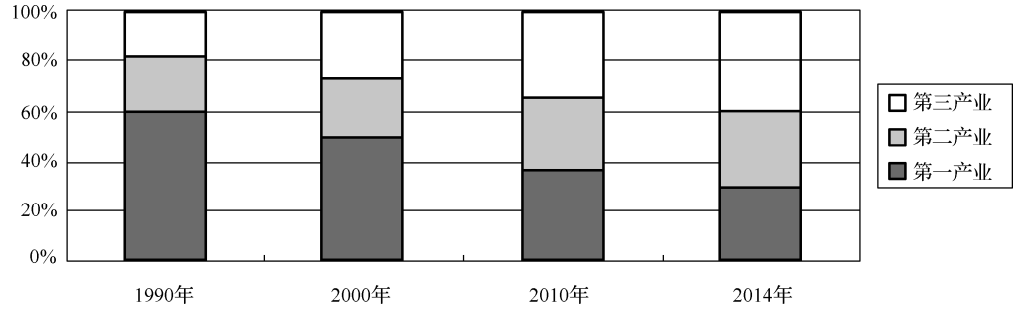


图 2.7　三大产业就业人员构成叠加式条形图

② 饼图。饼图也叫圆形图。作一个圆，把它分成若干个扇形，扇形面积大小与每一组频率相对应，圆形的面积主要用来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。适用饼图的情况：单组定性数据系列；分组数目一般不超过七个。

例如，表 2.12 体现了我国 2014 年客运量构成情况。

表 2.12　2014 年客运量构成

类　型	铁　路	公　路	水　运	民　航
数量(万人)	235 704	1 908 198	26 293	39 195
比重(%)	10.67	86.37	1.19	1.77

资料来源：《中国统计年鉴》(2015)。

根据表 2.12 所示资料制作的饼图如图 2.8 所示。

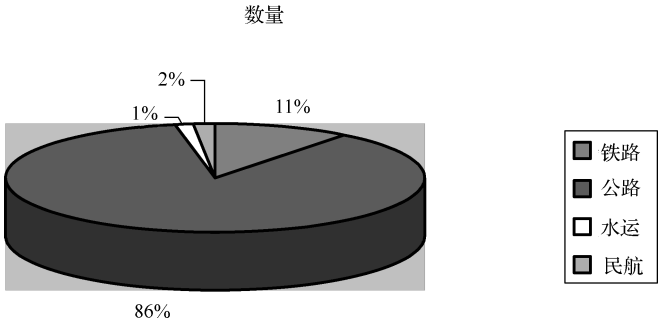


图 2.8　2014 年客运量构成饼图

③ 环形图。环形图由中间为“空洞”的多个同心圆组成，每一个总体占一个环，总体中每一部分的比例用环中的一段表示。环形图与圆形图一样可用于研究结构性问题。某届亚运会上中国、日本和韩国的奖牌数量如表 2.13 所示，通过 Excel 作环形图，如图 2.9 和图 2.10 所示。

表 2.13 某届亚运会上中国、日本和韩国的奖牌数量 单位：枚

国家 / 奖牌	中国	韩国	日本
金牌	147	56	30
银牌	71	47	55
铜牌	72	62	65
合计	290	165	150

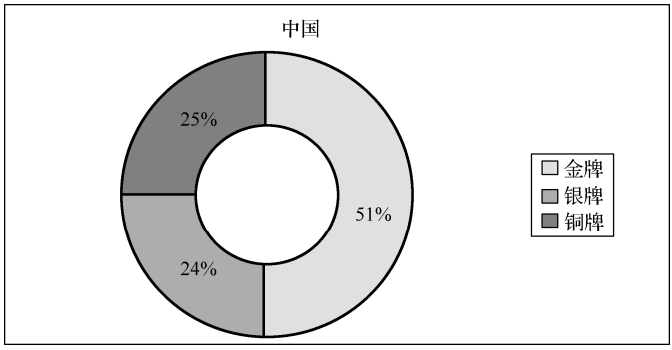


图 2.9 一个国家奖牌得数构成的环形图

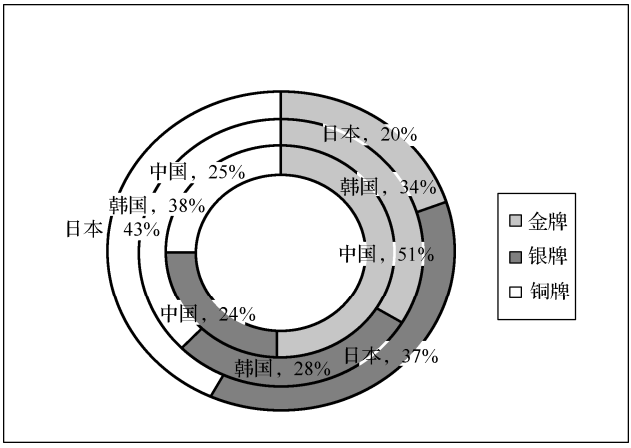


图 2.10 三个国家奖牌得数构成的对比环形图

④ 累计频数分布图。根据累计频数或累计频率，可以绘制累计频数或累计频率分布图。注意，定性数据中只有定序数据分组才能作累计频数或累计频率。“向上累计”就是从变量等级或强度低的向变量等级或强度高的方向把分布的次数依次累计相加，反之，则是“向下累计”。用表 2.9 中的学生成绩资料按优(90 分以上)、良(80～90 分)、中(70～80 分)、及格(60～70 分)、差(60 分以下)等级分类的累计次数绘制累计频数分布，如图 2.11 所示。

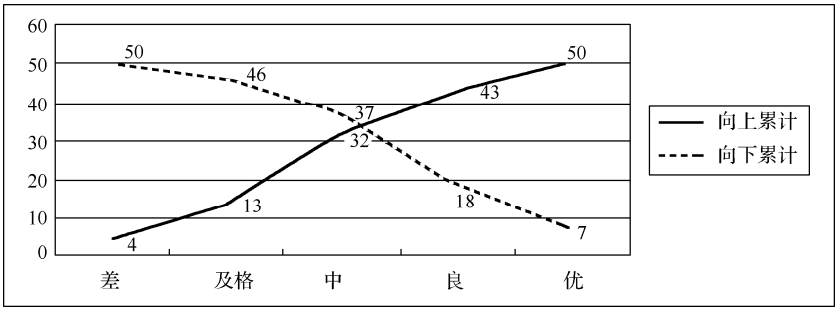


图 2.11　学生成绩累计频数分布图

(2) 定量数据的图形显示

① 直方图。直方图是用矩形的宽度和高度来表示频数分布特征的图形。在平面直角坐标中，横轴表示数据分组，纵轴表示频数或频率，各组与相应的频数就形成了一个矩形，即直方图。以前述例 2.1 中的数据和表 2.9 的整理结果为例绘制直方图，如图 2.12 所示。

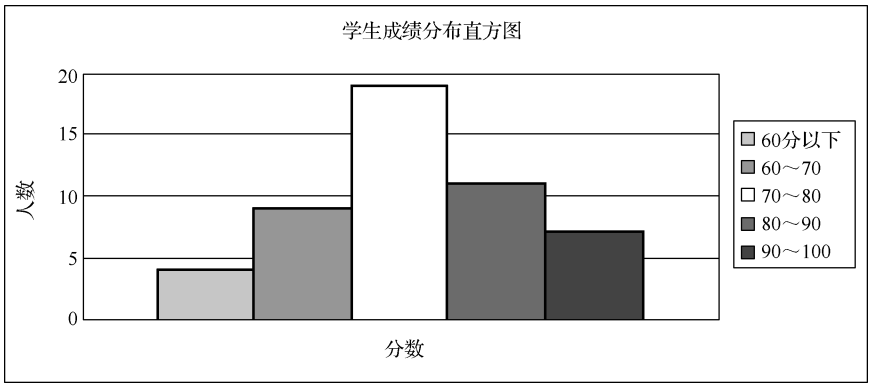


图 2.12　50 名学生统计学成绩分布直方图

对于等距数列，矩形的高度与各组的频数是成比例的，如果取矩形宽度(各组组距)为一个单位，用频率表示高度，则直方图下的总面积等于 1。

对于异距数列，为了更准确地反映总体的分布，纵轴应表示各组的频数密度。

频数密度=频数÷组距

② 折线图。折线图又称频数多边形图。在直方图的基础上，把各个矩形顶部的中点(组中值)连接起来，把原来的直方图抹掉就是折线图。考虑折线图与横轴所围成的面积要和直方图的面积相等，折线图两边端点要与横轴相交，具体作法：把第一个矩形的顶部中点和其竖边中点相连与横轴相交；把最后一个矩形的顶部中点和其竖边中点相连与横轴相交。根据例 2.1 中的数据和表 2.9 的整理结果为例绘制折线图，如图 2.13 所示。

③ 曲线图。对于折线图来说，组数越多或者组距越小，则折线越光滑，理论上，当组数趋于无限多或组距趋于无限小时，折线就成了曲线。曲线图画法与折线图画法一样，只是在连接各点时不用折线而用平滑的曲线，如图 2.14 所示。

④ 茎叶图。茎叶图可以看成是带数据的直方图。前述的直方图、折线图和曲线图虽能直观形象反映一组数据的频数分布状况，但是经过了分组整理，损失了原始数据信息。茎叶图是将传统的统计分组和绘制直方图工作一次完成，既反映数据分布特征，又保留了每一个原