第一章
统计学概述

学习目标

了解统计学的研究对象；

掌握统计学的几个基本概念；

了解统计学的分支；

掌握统计数据的类型；

了解统计数据的收集方法；

理解统计学的正确应用。

引　例

“统计”这个词大家可能并不陌生，翻开报纸杂志，到处充斥着调查数据和用统计方法研究所得到的结论。在 Google 上检索“统计分析表明”一词，有很多结果出现，现摘引若干如下。

(1) 中国人民银行对我国 2002 年 2 月企业商品价格统计分析表明：我国企业商品价格总水平较上月上升 0.1%，较上年同期下降 2.7%，同比降幅较上月缩小 0.3 个百分点。

(2) 中华人民共和国国家知识产权局对我国 19 个省、50 个城市 2005—2007 年授权专利的个人、大专院校、科研单位和企业的专利实施状况进行了调查。统计分析表明：86%已经收回研发投入的成本，71.5%获得的收益大于研发投入。

(3) 全国 30 个省、自治区、直辖市消费者协会统计分析表明：2007 年共受理消费者投诉 656 863 件，为消费者挽回经济损失 83 964 万元。

(4) 2009 年就业蓝皮书《2009 年中国大学生就业报告》的统计分析表明：本科与高职高专的毕业生对雇主的满意程度分别为 70%和 68%，其中“工作要求与压力”不满意度最低，“薪资福利”和“个人发展空间”不满意度最高。

(5) 统计分析表明：在总体的社会化方面，流动儿童优于留守儿童，流动儿童社会化均值分高出留守儿童 1.819 个百分点，差异达到统计显著水平。

(6) 统计分析表明：在父母经常吵架的家庭中，孩子的心理问题检出率为 31.68%，离婚

家庭的为 30.30%，和睦家庭的为 18.88%。

(7) 气象卫星遥感监测统计分析表明：2010 年 1 月 23 日前后，渤海全海域及辽东湾、渤海湾、莱州湾海冰面积均达 2000 年以来同期最大值，其中，莱州湾海冰面积较常年同期平均值偏大约 5.6 倍。

(8) 统计分析表明：胎次对母猪产仔总数具有影响，其中，1、2 胎差异显著($P<0.05$)，1、3 胎差异极显著($P<0.01$)，其他各组之间差异均不显著。

……

由上述资料至少可以激发以下四点思考。

第一，这些资料覆盖了经济、社会、科技、心理、气象等各个方面，由此统计应用范围之广可见一斑。

第二，对比不同的资料可以看到，有的统计分析是基于对基层数据的汇总，如资料(2)和资料(3)；有的来自专门组织的调查，如资料(1)、资料(4)、资料(5)和资料(6)；有的来自设备监测，如资料(7)；有的则来自科学实验，如资料(8)。由此可见，统计分析并不受数据来源的局限。

第三，前面七项资料对统计分析结果的陈述并无难理解之处，而第八项资料中则出现了一个符号 P，而且资料(5)和资料(8)都用了一个看似普通，实则体现统计思想的术语——显著。

第四，以上资料大都指出“统计分析表明”，那么它们分别使用了哪种统计分析方法呢？

实际上，统计学的思想和方法远比上述资料所显示的深邃，其符号和术语也远比资料中用到的丰富。在统计学的学习起点，不妨先对统计学的基本内容和基本原理作一个剪影式的了解。

第一节　统计学及统计方法的基本思想

一、统计学及其若干基本概念

从字面来看，统计就是统而计之，即将个别数据综合起来得到结论。虽然统计活动是不分国界、人类早已有之的活动，但是对数据进行系统的描述，并在此基础上进行推断的科学原理和操作方法，即作为一个学科门类的统计学，却是一门舶来的学问。《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。”这一定义言简意赅，其中蕴涵了以下几层含义。

(1) 统计学的研究对象是数据。

(2) 统计学是一个围绕数据的全过程研究，该过程始于数据的获取，经过对数据的分析，最终从数据中提取信息并得出结论。

(3) 统计学是一门“硬”科学，因为统计方法具有坚实的数理基础。也有人认为统计学兼具艺术性，这是针对统计应用而言的。在分析实际数据时，需要灵活使用统计方法。实际

中，常常会遇到这样的情形：对于同样的数据，不同的分析者使用不同的统计分析方法，得到不同的结论。正是由于统计应用具有灵活性，才需要强调正确应用统计方法，避免对统计方法的误用与滥用。关于这一点，在第三节中有专门的论述。

统计涉及两对基本概念：一是总体和样本，二是参数和统计量。

所谓总体，是指研究所关注的全部单元组成的集合。例如，如果一个研究者希望了解某地区住户的收入，则该地区的全部住户就是总体。如果总体包含的单元很多，受时间和经费等条件的限制，往往不能对总体中的所有单元都进行调查，而只能抽取总体中的一部分单元进行调查。例如，若某地区的住户特别多，逐一进行调查的成本很高，此时可以考虑从中抽取一部分住户进行调查。还有些时候，为了获取所需数据，需要进行破坏性实验，为了减少损失，不必对总体中的每个单元都进行实验，而是抽取一部分单元进行实验。例如，为了确定某人的血铅含量是否超标，只需要抽取他的少量血液进行测量即可。这些从总体中抽取出来的单元所组成的集合就称为样本，而样本中所含单元的数目则称为样本容量。

对总体和样本的另外一种理解是将它们对应于研究所关注的具体特征。例如，在前面的例子中，总体为某地区住户的收入，样本则为抽出来接受调查的住户的收入。

在实际研究中，研究者常常需要了解总体的一些经过汇总后的数据特征，而不是每一个总体单元的具体特征。例如，在前面的例子中，研究者或许希望了解的是某地区全体住户的户均收入，即将全体住户的收入加总之后再除以全部住户数目所得到的数据。这种特征是总体的数学期望，在后面的章节中有专门介绍。同样，对样本也可以计算相应的数据特征。如果计算某地区样本住户的户均收入，则为样本均值。这种对总体数据加工出来的数据特征称为总体参数，对样本数据加工出来的数据特征称为样本统计量。

需要注意的是，研究者真正感兴趣的是总体或总体参数，而不是样本或样本统计量。如果数据只能在样本范围内获得，则对样本数据进行分析（描述统计）之后，研究并没有结束，还需要根据样本的分析结果，在一定的方法支持下对总体或总体参数进行推理和判断（统计推断）。例如，研究者需要依据样本住户的户均收入水平推断总体住户的户均收入水平。

由于样本只是总体的一部分，所以统计推断结果难免会有误差。为了控制推断误差，需要抽取对总体有较强代表性的样本。

二、统计方法的基本思想

统计方法是实证分析中收集和分析数据的重要工具，几乎所有科学都要运用统计方法。但是，在学习和应用统计方法的同时，一定要认识到“统计学不止是一种方法或技术，还含有世界观的成分——它是看待世界上万事万物的一种方法[①]”。

事实上，在统计模型成为科学研究的范式之前，科学界奉行的是一种固化的哲学观，即机械式宇宙观。这种哲学观认为，所有的物体都按照一定的规律运动，所有未来的事件都取

① 陈希孺. 数理统计学简史[M]. 长沙：湖南教育出版社，2002.

决于过去的事件。按照这种观点，用少数几个公式就可以描述现实世界的一切，而且只要有一套完整的公式和足够精确的数据，就可以对未来事件进行预测。

然而，随着科学的发展，人们发现无论是自然科学研究还是社会科学研究，都不可避免地存在误差与不确定性，因此任何对现实世界的描述以及对未来的预测都只能是一种逼近，其所能达到的最好的研究结果只能是给出现实状态与未来可能结果的概率分布，这种概率分布就是统计模型。参照统计模型，人们可以对不确定性有一个定量的把握，并据此作出各种决策。例如，两个企业生产同类产品，其次品率分别为 5%和 2%。这就是一个简单的概率分布，如表 1-1 所示。

表 1-1 两个企业产品质量的概率分布

企 业	次 品 率	合格品率
企业 1	0.05	0.95
企业 2	0.02	0.98

显然，由于生产设备和生产环境的复杂性，没有哪个企业能够确定无疑地总是生产合格品。如果没有产品质量的概率分布，购买者将犹豫不决。然而，通过对比两个企业产品质量的概率分布，购买者很容易作出正确决策。当然，现实世界中的决策远不止这么简单，还要考虑各种因素。例如，在上面的例子中，购买者可能还需要考虑两个企业产品的价格和运输成本等，但毋庸置疑的是，产品质量的概率分布依然是购买决策中的关键信息。

一般来说，概率分布是科学研究中可以获得的最为全面的信息，研究者可以基于概率分布计算出其所感兴趣的任何参数。然而概率分布并不是总能获得的，此时，研究者会退而求其次，转向对关键参数的研究。不过即使是对关键参数的研究，依然离不开概率分布。仍以对某地区住户收入的研究为例，最全面的信息莫过于住户收入的概率分布，掌握了这个概率分布，研究者可以知晓该地区住户收入的各种特征，如中位数收入、贫困住户比重以及收入不均等程度等。如果该概率分布不可得，则研究者可能希望了解该地区住户收入的关键特征，如平均收入，并依据对样本平均收入的概率分布的假定推断出总体的平均收入。

三、统计学的大家族以及统计学与其他学科的关系

（一）统计学的大家族

自 17 世纪中叶一批数学家对概率的数学理论进行研究以来，经过三个多世纪的发展，统计学已经形成一个建立在数理统计学原理基础之上、集聚各种统计方法的庞大家族。限于篇幅和写作目的，本书不会也不可能涉及所有的数理统计原理和统计分析方法。下面对统计学的大家族进行简单介绍，以使读者更加清楚地认识统计学的功能。

（1）根据统计分析的阶段不同，统计可以划分为描述统计和推断统计两个分支。相应地，统计学可以分为描述统计学和推断统计学。可以说，描述统计是推断统计的基础，推断统计是描述统计的高级阶段。两者的根本区别在于，描述统计是对确定的样本数据的分析，没有不确定性；而推断统计则是依据样本数据对总体特征进行推断，具有不确定性，需要借助概率这一工具。

描述统计和推断统计的关系如图 1-1 所示。

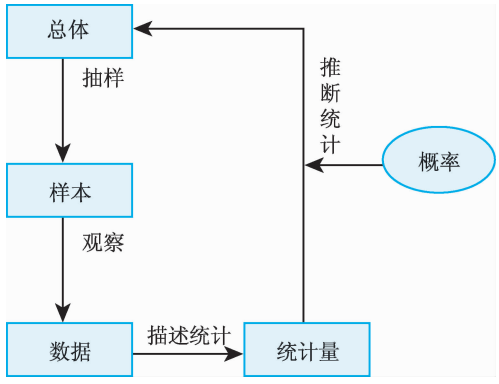

图 1-1　描述统计与推断统计

（2）根据研究的内容不同，统计学可以划分为理论统计学和应用统计学两个分支。理论统计学在国外又称为数理统计学，其研究内容是统计学的概率，可视为一门纯数学。应用统计学则是在理论统计学的基础上对具体统计方法的研究。在应用统计学中，根据是否假定总体的概率分布只依赖于有限个实参数，又可以分为参数统计方法和非参数统计方法。前者有此假定，而后者则没有这个假定。

（3）根据对概率和参数的理解不同，统计学可以划分为频率学派和贝叶斯学派两个分支。频率学派理解的概率是频率概念，而贝叶斯学派理解的概率是主观概念；频率学派认为参数是一个客观存在的常数，而贝叶斯学派则认为参数也是一个可以用概率分布刻画的随机变量。

总体来说，本书的内容包括基础的数理统计和常见的应用统计方法，主要介绍频率学派观点，以参数统计方法为主。如果读者对贝叶斯学派和现代非参数统计方法感兴趣，可以参见相关文献。

（二）统计学与其他学科的关系

统计学在发展过程中，通过将统计方法应用于其他学科的研究，与其他学科保持了密切的联系。“在 20 世纪之前本无‘专职’的数理统计学家，统计学家都是某一专门学科领域的专家，因工作上的需要研究数据分析问题而介入统计学。”[①]例如，对统计学的发展作出重大贡献的罗纳德·爱尔默·费希尔(R. A. Fisher)，“在遗传学方面的名声不亚于统计方面，他的研究论文不少发表在《优生学杂志》”[②]。

今天，统计学继续保持着与其他学科交叉发展的关系，而且应用领域更加广泛。一方面，某些学科在解决本领域的数据分析问题时，借助于统计方法，产生出特定的交叉学科。例如，计量经济学、金融计量学、历史计量学、文献计量学等，都是此类学科。另一方面，自然科学和社会科学的各个方面的实证研究都需要应用统计方法。例如，近年来统计分析在生物医学、金融资产定价、质量管理过程等领域的应用均取得了丰硕的成果。

① 陈希孺. 数理统计学简史[M]. 长沙：湖南教育出版社，2002：270.

② 陈希孺. 数理统计学简史[M]. 长沙：湖南教育出版社，2002：270.

专栏 1-1

信息时代统计学面临的挑战①

与费希尔时代的统计学家相比，当代统计学家最大的优势在于掌握了强大的计算技术以及由此带来的海量数据。为了统计学的存续和发展，统计学家必须将解决实际的数据问题作为自己的主要目标，而且统计学的发展必须跟上信息技术的发展。

为了达到这一目标，统计学家必须面对许多挑战，如顺应网格计算趋势、有效利用数据库及其他数据源（如传感器网络），使用新的或非传统的数学结果，发展新的满足通信和计算能力约束的统计算法，以及设计出能够兼容上述努力的新的统计推断范式。

在组织或文化的层面，统计学界也面临诸多挑战。在许多大型的科学计划中，如大气科学（如模型模拟和遥感数据）、天文学（如数字巡天）以及生物学（如基因或脑切片扫描数据库）等，收集和管理着海量的数据。其中虽然可以发现单个统计学家的身影，但单个统计学家很难影响到这些计划的基础。例如，他们难以在收集数据和挖掘大型数据库时的算法选择等问题上拥有发言权。如果统计学想要在信息时代产生必要的影响，就需要进行集体思考，树立统计学界的领导力。

除了统计技术，要想与科学家们合作成功，并说服科学家们承认统计学在科学研究中的重要作用，还需要高超的社交技巧。这些专业外的技巧在跨学科研究中的重要性表明，统计学界需要转变文化。统计学家应当珍视这些非传统技巧，并承认它们在诸如终身教职审查、职称提升以及获奖等方面的作用。

最后，但不是最次要的，统计学家需要将相关的专业知识和社交技巧传授给本专业的研究生和本科生。

第二节 统 计 数 据

一、统计数据的概念和特征

统计学的研究对象为数据。所谓数据，是指对研究对象的某种特征进行测量的结果。要想正确理解统计数据，应当注意统计数据的以下两个特征。

（1）统计数据的表现形式是多样的，既可以是数字，也可以是文字。而且，研究者根据需要，在对同一个对象进行测量时，可以灵活选择数据的形式。例如，对住户的收入进行测

① Bin Yu. Embracing Statistical Challenges in the Information Technology Age[J]. Technometrics. Vol. 49. No. 3. (August. ,2007):pp. 246-247.

量，其结果可以表现为具体的数字，如每月××元；也可以是收入水平的一个文字刻画，如高、中、低。数据的表现形式是数据的测量尺度问题，这在下文中有专门的介绍。

（2）统计数据是带有随机性的，而不是确定的。所谓随机性，是指数据可以通过某种概率分布规律来描述。这就将统计学与其他处理数据的学科（如数值分析）区别开来。需要注意的是，数据的随机性是针对样本的理论取值而言的，而不是针对样本的具体取值而言的，后者一经测量，便可作为确定的数值。例如，抽取一部分住户进行观察，以研究某地区住户的平均收入。由于事先不能确定会抽出哪些住户，所以从理论上说，样本数据是随机的；然而，一旦抽出一个具体的样本，则被抽中住户的收入数据就是确定的。

二、统计数据的类型

从不同的角度可以将统计数据划分为不同的类型。

1. 根据测量尺度分类

根据数据的测量尺度（即精度），可以将数据分为定性数据和定量数据。

定性数据是指以文字形式表现的数据。例如，性别的测量结果为男或女，满意度的测量结果为非常满意、比较满意、一般、比较不满意或者非常不满意。根据是否可以对数据进行排序，定性数据又分为两类：不能排序的是定类数据（categorical data，又称nominal data），可以排序的是定序数据（rank data，又称 ordinal data）。显然，性别数据属于定类数据，而满意度数据则为定序数据。

为了方便数据分析，通常需要将定性数据转化为数字形式，这就是编码（coding）。例如，对性别进行测量时，用 1 表示男性，用 0 表示女性；对满意度进行测量时，用 5 表示非常满意，用 1 表示非常不满意，用 2、3 和 4 依次表示中间的满意程度。编码之后，定性数据就可以参加各种数学运算了。但一定要注意，这种编码数字的本质仍然是文字，不同于真正的数字。例如，性别的编码不能比较大小，相加没有意义；满意度的编码只能比较大小，相加同样没有意义。正因为如此，在进行统计分析时，对定性数据的编码要注意采用正确的方法，否则会出现不符合实际的分析结果。

定量数据是指以数字形式表现的数据。例如，考试成绩的测量结果为××分，收入的测量结果为××元，公路里程的测量结果为××千米，股票价格指数的测量结果为××点等。

显然，从测量精度或者从数据所包含的信息量来讲，定量数据要高（或多）于定性数据，而定序数据又高（或多）于定类数据。此外，测量尺度较高的数据可以转化为测量尺度较低的数据；反之，则行不通。例如，可以将考试成绩的定量数据转化为定序数据（如 98 分为优，65 分为及格等），反过来却无法判断一个成绩为优的考生究竟考了多少分。因此，在收集数据时，要根据所测量对象的特点，尽量在较高级的尺度上测量，这样可以保留尽可能多的信息，也便于数据类型的转化。

2. 根据数据收集方法分类

根据数据的收集方法，可以将统计数据分为实验数据和观察数据。

实验数据是指在实验之前并不存在，需要通过事先的实验设计，在实验中控制实验对象而收集的数据。例如，在研究杀虫剂的剂量对虫子死亡的概率的影响时，会有严格的实验设

计，在此实验中收集的数据（包括杀虫剂剂量和虫子的生存状态等）就属于实验数据。一般来说，自然科学研究的数据多为实验数据。

观察数据是指客观上已经存在，但是需要经过观察或询问才能获得的数据。这类数据通常需要通过抽样调查来收集。例如，设计一个抽样调查方案，对某地区住户进行抽样调查，调查所收集到的住户信息（包括收入、家庭人口、户主职业等）就属于观察数据。一般来说，社会科学研究的数据多为观察数据。

在统计学中，因果关系推断是一个非常重要的研究目标。需要指出的是，在推断因果关系时，由于实验设计可以更好地控制实验环境和实验条件，从而更好地解决其他因素的干扰，所以实验数据要优于观察数据。

3. 根据数据结构分类

根据数据的结构，可以将数据分为截面数据和历时数据。

截面数据是指在同一时点或同一时期，就多个个体收集的数据。例如，某地区 500 名住户在 2009 年的年收入就属于截面数据。

历时数据是指在多个时点或多个时期，就一个或多个个体收集的数据。例如，某企业历年的净利润数据就属于历时数据。如果是对固定的多个个体收集的历时数据，则称为面板数据。例如，2000—2010 年，某地区固定的 500 名住户的相关数据就属于面板数据。

由于因果关系推断的一个基本准则是时间顺序，即因在前、果在后，所以在因果关系推断中，历时数据要优于截面数据，其中面板数据是更为理想的数据结构。

三、统计数据的收集

（一）实验设计

实验设计是指对实验进行科学合理的安排，以达到最好的实验效果。一个科学的实验设计，能够合理地安排各种实验因素，严格地控制实验误差，从而获取有效的实验数据，为统计分析提供支持。

实验设计的基本步骤如下。

（1）随机选择实验对象。

（2）将实验对象随机分为两组：一组接受实验处理，即实验组；另一组不接受任何处理或接受一些对照处理[①]，即对照组。

（3）前测，即在实验前收集两组实验对象的有关数据。

（4）进行实验。

（5）后测，即在实验结束后再次收集两组实验对象的有关数据。

（6）分析前测和后测的数据，得出结论。

实验设计示意图如图 1-2 所示。

① 常用的对照处理包括安慰剂对照、实验条件对照、标准对照以及历史对照等。

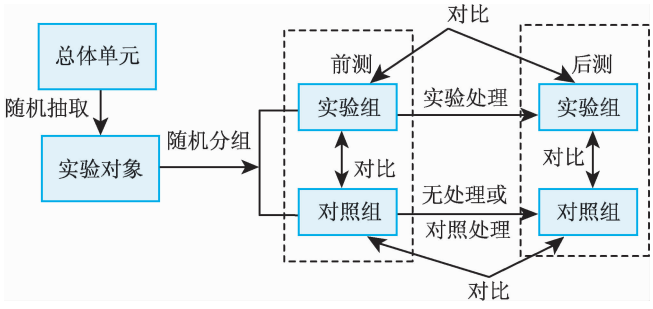

图 1-2　实验设计示意图

例如，要检验某种减肥药是否有效，有 n 名肥胖症患者同意接受实验。按照实验设计的要求，将 n 名患者随机分为两组，并测量每组患者的体重。尽量保证两组患者的前测结果是没有显著差异的。实验组患者服用该减肥药，对照组患者则服用外观相同的安慰剂①。一个疗程结束后，再次测量两组患者的体重。如果在后测中，两组患者出现了显著差异，则可以认为是减肥药产生的效果。

在实验设计中，设置实验组和对照组、将每个实验对象随机分入实验组或对照组、进行两次测量这三个要素非常重要。如果一个实验设计违背了随机化原则或对照原则或者没有进行前测，只能称为准实验设计。利用准实验数据进行分析，会面临更大的困难。

（二）抽样调查

抽样调查是指从总体中抽取一部分单元作为样本，根据对所抽取的样本进行调查，获得对总体的了解。按照抽样时是否遵循随机原则，即总体的每个单元是否都有非零的入样概率，抽样调查可分为概率抽样和非概率抽样两类。其中，概率抽样以随机原则抽取样本，非概率抽样则不然。一般而言，在推断统计中应使用概率抽样收集的数据。下面介绍几种基本的概率抽样方法。

1. 简单随机抽样

简单随机抽样（simple random sampling）是一种最基本的抽样方法，是指从抽样总体②的 N 个单元中随机地、一个一个地抽取 n 个单元作为样本。简单随机抽样具有简单直观、便于统计推断的优点，经典的统计推断理论——假定数据，就来自简单随机抽样。但是该方法要求有总体单元的名单，且入样单元比较分散，因而给调查带来一定的难度。通常大型调查很少直接采用简单随机抽样，而是将这种方法与其他抽样方法结合起来使用。

2. 分层抽样

分层抽样（stratified sampling）是指将抽样单元按照某种特征划分为不同的层，然后从各层中独立、随机地抽取样本。分层抽样保证了样本的结构与总体较为相似，有助于提高统

① 严格的实验设计还要做到令对照组患者不知道自己服用的是安慰剂。

② 抽样总体是指从中抽取样本的总体，与目标总体可能有出入。

计推断的精度。此外，分层在一定程度上方便了调查的组织和实施，因此分层抽样在实践中被广泛应用。

3. 整群抽样

整群抽样(cluster sampling)是指将总体中若干单元合并为群，抽样时直接抽取群，然后对抽中的群所包含的各个单元进行调查。整群抽样的最大优点在于不需要有总体单元的名单，且群通常都是按地理位置划分的，因而极大地减小了入样单元的分散性，有利于调查的组织实施。但是，由于群内单元的相似性较高，因此利用整群抽样的数据进行统计推断时，推断的精度较低。

4. 系统抽样

系统抽样(systematic sampling)是指将总体单元按照一定的顺序排列，先随机抽取一个样本单元，然后按照事先规定好的规则抽取其他样本单元。系统抽样虽然操作简便，但是对统计量方差的估计比较困难，不利于统计推断。

5. 多阶段抽样

两阶段抽样与整群抽样有些相似，第一阶段抽取群，但接下来并不是对群内所有单元进行调查，而是抽取若干个单元进行调查。如果第二阶段继续抽取群，接下来再抽取若干单元进行调查，依此类推，便形成了多阶段抽样(multi-stage sampling)。通常，大型调查大都采用多阶段抽样。

第三节 正确认识和使用统计方法

一、正确认识统计方法

(一) 统计方法的中立性

虽然统计方法广泛应用于各个领域的研究，但它只是一种中立的研究工具。在具体的应用中，统计方法不坚持任何学科中的任何观点。换句话说，统计方法只回答“是什么”的问题，而不回答“应该是什么”的问题。如果有人不同意使用统计方法，完全可以不用它，只作纯粹定性的讨论。但是，只要想进行实证分析，就必须按照统计学的规范来收集和分析数据，由数据来揭示结论。例如，20 世纪六七十年代，美国政界关于对付犯罪的根本途径分歧严重，自由派认为入狱刑罚具有破坏性，不利于改造罪犯、降低犯罪率，有些专家甚至认为某些犯罪行为不应该被看做犯罪。这些观点都只是定性的讨论，现实中入狱刑罚与犯罪之间究竟呈何种关系，需要进行客观的检验。一些犯罪学家的统计研究结果显示，入狱刑罚能够大大降低犯罪率。史蒂夫·莱维特(Steve Levitt)通过研究监狱诉讼的结果发现，从监狱里每释放一名犯人，每年的犯罪数量就会增加 15 起。这些实证研究的结论使得美国政界在审判政策和政策实施方面达成了共识，自由派不再像 20 世纪六七十年代那样抵制入狱刑

罚了。

（二）统计结论的表层性

统计分析的结果是对表层数量关系的揭示，完全不触及问题的专业内涵，因此如果没有专业理论的支持，不可随意将统计分析结果推定为因果关系，即统计方法只回答“什么与什么有关”，而不回答“什么决定什么”的问题。例如，如果收集某地区各月冷饮消费量与溺水死亡人数的数据，经统计的相关分析可以发现两者具有非常强的正相关性，但是并不能据此判断冷饮是造成溺水的原因。真正的原因可能是：天气炎热导致冷饮消费和游泳人数增加，后者伴随着溺水死亡人数的增加。有时真正的原因并不这么明显或者相关关系“非常像”因果关系，导致研究者在分析中容易犯将相关关系推定为因果关系的错误，因此要切记统计结论的表层性特点。

（三）统计方法与统计软件

在计算机日益普及的时代，统计软件获得了极大的开发，各种功能强大、界面友好的统计软件层出不穷。客观地说，统计软件的发展在很大程度上推动了非统计专业的研究者对统计方法的使用。一个很容易被忽略的事实是：能够熟练使用统计软件并不等于能够正确应用统计方法。举一个简单的例子，如果分析中涉及定性数据，而研究者对定性数据的编码不正确，则统计软件就会根据编码错误的数据输出结果，其结果自然不可取，但是计算机并不会提醒用户结果不正确。只有研究者凭借自己的专业知识才能发现编码有误，通过重新编码从根本上解决问题。因此，学习统计方法一定要知其然、知其所以然，而不能完全依赖统计软件。

二、误用统计方法的两个例子

统计方法在各个领域都有应用，尤其是在实证主义盛行的今天，统计数据和统计分析更是大显身手。统计数据与统计分析的流行昭示了人类科学意识的加强，但使用者也要充分意识到统计方法并非“灵丹妙药”，如果使用不规范，不仅不能达到期望的目标，甚至可能造成严重的后果。下面列举两个误用统计方法的例子，以期引起初学者的注意。

（一）样本有偏造成的分析偏差

正确的分析结果建立在对总体具有代表性的样本的基础上，如果样本有偏，则很容易造成分析偏差。假定两个变量之间的真实关系如图 1-3 所示，显然两者不存在相关关系。现在通过抽样调查来发掘两者的关系。如果研究者有意得到两者有相关关系的结论，故意收集了一个有偏的样本，即图 1-3 中由椭圆围起来的单元，则相关关系似乎是成立的，而真实的关系就在看似客观的统计分析之后被隐藏起来了。由此可见，为保证分析结果的有效性，对数据收集过程和数据质量无论怎样强调都不为过。

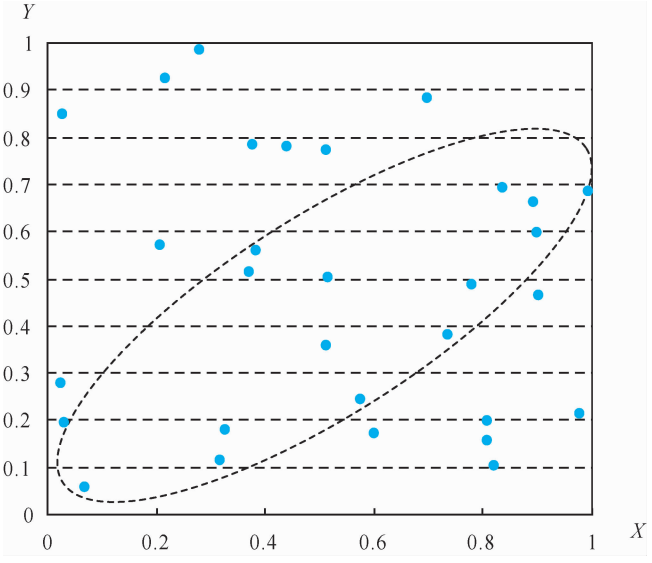

图 1-3　样本有偏造成的分析偏差

（二）错误的图示方法造成的分析偏差

统计图是探索数据规律的基本工具。统计图绘制得好，能够达到“一图胜千言”的效果。但是，如果研究者出于某种目的有意对统计图进行歪曲，则很容易造成分析偏差。在图 1-4 的三幅图中，直观地看，(b)反映出来的波动幅度最小，(c)反映出来的波动幅度最大，(a)则介乎两者之间。但实际上，它们刻画的是同一个变量的演变轨迹，只不过图的坐标比例不同而已。显然，简单地拖拉鼠标，操控图的形式，就可以“支持”某个特定的观点。

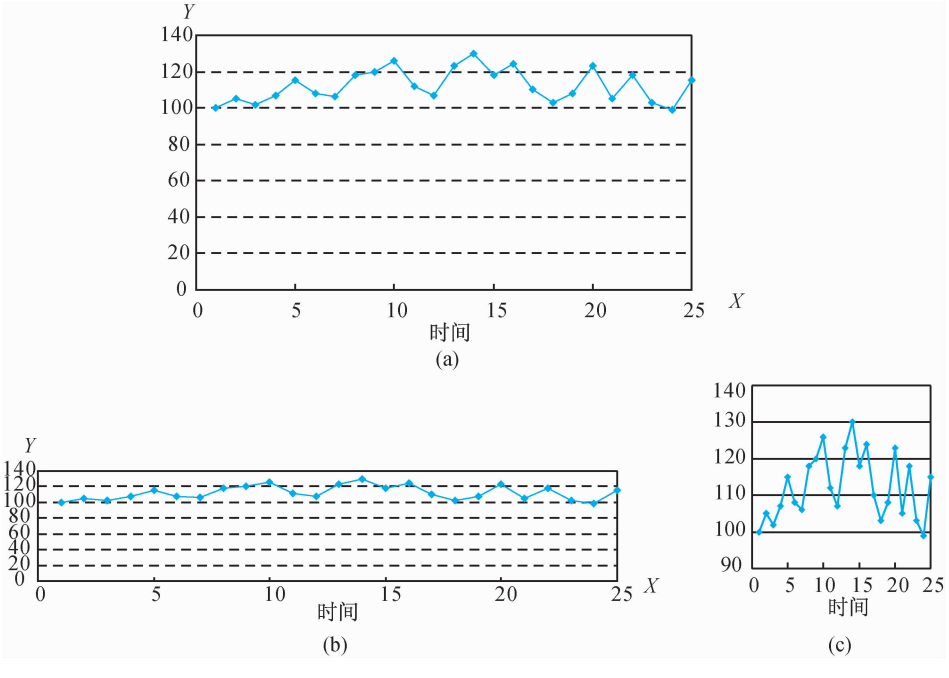

图 1-4　错误的图示方法造成的分析偏差

由上述两个例子可以清楚地看到，如果不遵照统计分析的规范，得出的结论就会出现偏差。对于复杂统计方法的应用而言，陷阱会更多。因此，在应用统计方法时，一定要保持谨慎的态度，避免走进误区。

主要术语

统计学　总体　样本　参数　统计量　描述统计　推断统计　统计数据　定性数据　定类数据　定序数据　定量数据　实验数据　观察数据　截面数据　历时数据　实验设计　抽样调查

思 考 题

(1) 统计学的研究对象是什么？
(2) 参数和统计量的关系是什么？
(3) 统计方法的基本思想是什么？
(4) 统计数据的特征是什么？
(5) 统计数据的类型有哪些？
(6) 如何理解统计方法的中立性？
(7) 找一个应用统计方法进行分析的例子，看其分析过程有无错误。

第二章
统计数据的描述

学习目标

掌握统计数据频数分布的构建；

掌握统计数据的图示方法；

掌握统计数据的集中趋势和离散程度的测量方法；

熟悉统计数据偏度和峰度的测量方法。

引　例

东北地区是我国重要的商品粮基地。随着全球粮食危机的出现，东北地区的自然环境及其变化和对区域农业生产的影响已成为我国环境科学研究的热点。研究人员基于东北地区百余年来月降水量的实测资料，对其降水量变化特征和统计规律进行了分析。该研究提供了哈尔滨、长春和沈阳三个城市1906—2007年降水量的统计特征，如表2-1所示。

表2-1　哈尔滨、长春和沈阳1906—2007年降水量统计特征①

城　　市	均值/mm	离散系数	最小值/mm	最大值/mm	峰度系数	偏度系数
哈尔滨	544.7	0.22	342.8	1 041	1.96	0.8
长春	604.0	0.21	329.7	970.5	0.3	0.57
沈阳	709.8	0.21	341.1	1 064.9	−0.3	0.31

根据表2-1中的数据，关于东北地区的降水量能得到哪些结论？表2-1中出现的最小值和最大值不难理解，但均值、离散系数、峰度系数和偏度系数等的含义是什么？学习本章内容之后，这些问题将迎刃而解。

① 高鹏，穆兴民，王飞，等. 中国东北地区近百年来降水量变化趋势分析[J]. 水文，2010(5)：18，80-84.

描述统计是统计分析的第一步，本章将介绍如何对样本数据进行描述性统计分析，具体包括频数分布的构建、统计图示方法的使用以及几种分布特征统计量的计算。

第一节　频数分布

原始数据是在实验设计或统计调查中就个体收集的数据，如每个人的性别、职业以及收入等。原始数据没有经过任何处理，又称为未分组数据。如果数据量少，分析者可以通过观察原始数据来把握数据的结构。但当数据量很大时，分析者就必须对原始数据进行某种整理才能发掘数据的内在结构，而分组是数据整理的主要内容。经过分组的数据称为分组数据。将分组数据和未分组数据区别开来是非常重要的，因为用于计算两种数据的统计量的方法是不同的。

频数分布是一种用来对数据进行分组的非常有效的工具，它以组和频数的表格形式对数据进行总结。对于定性数据而言，构建频数分布比较容易；对于定量数据来说，构建频数分布则比较困难。

一、定性数据频数分布的构建

由于定性数据的取值有限，所以对于定性数据而言，其每一个取值就可以作为一组。只要知道每个取值出现的次数，即频数，并编制频数分布表，就完成了定性数据频数分布的构建。

【例 2-1】　某学校为了研究学生对图书馆的满意度，调查了 30 名学生。被调查学生的性别和满意度数据如表 2-2 所示。

表 2-2　学生对图书馆满意度调查结果

编　号	性　别	满意度	编　号	性　别	满意度	编　号	性　别	满意度
1	1	3	11	0	3	21	1	2
2	1	4	12	0	4	22	0	3
3	1	3	13	1	5	23	1	4
4	0	5	14	1	5	24	1	5
5	0	5	15	1	4	25	0	3
6	0	5	16	1	3	26	0	1
7	1	5	17	1	2	27	0	3
8	1	2	18	1	2	28	1	1
9	0	1	19	0	4	29	1	4
10	0	3	20	0	3	30	1	2

注：对于性别，1 表示男生，0 表示女生；对于满意度，5 表示非常满意，4 表示比较满意，3 表示一般，2 表示比较不满意，1 表示非常不满意。

构建上述数据的频数分布。

解　对性别和满意度分别按照其取值进行分组，其中，性别分为两组，即男生组和女生

组；满意度分为五组，即非常满意组、比较满意组、一般组、比较不满意组和非常不满意组。进一步整理出其频数分布，如表 2-3 和表 2-4 所示。

表 2-3 性别的频数分布

组	频 数
男生	17
女生	13
合计	30

表 2-4 满意度的频数分布

组	频 数
非常满意	7
比较满意	6
一般	9
比较不满意	5
非常不满意	3
合计	30

从例 2-1 中容易看到，直接观察表 2-2，数据比较凌乱，难以把握被调查者的性别结构和满意度分布状况，而表 2-3 和表 2-4 则能够清晰地显示性别和满意度的分布。根据表 2-3，可以得知被调查者中男生比女生多 4 名。根据表 2-4，可以得知对图书馆感觉一般的学生最多，约占总数的 1/3，而且对图书馆感到满意的学生要多于对图书馆感到不满意的学生，前者有 13 名，后者只有 8 名。

关于定性数据的频数分布，需要指出以下两点。

（1）定类数据的取值没有顺序，因此频数分布表中组的排列顺序是随意的。例如，表 2-3 中的性别顺序可以随意排列。而定序数据的取值有顺序，因此频数分布表中的组要按照一定的大小顺序排列，当然既可以升序排列，也可以降序排列。例如，表 2-4 中的满意度顺序还可以颠倒过来，即从非常不满意到非常满意。

针对定性数据能够排序的性质，可以在频数分布的基础上构建累积频数分布。所谓累积频数，是指按照某种顺序，将各组的频数逐级累积起来得到的频数。例如，基于表 2-4 可以得到表 2-5 所示的累积频数分布。

表 2-5 满意度的累积频数分布

组	频 数	累积频数
非常满意	7	7
比较满意	6	13
一般	9	22
比较不满意	5	27
非常不满意	3	30
合计	30	—

（2）有时，定性数据的取值较多，如职业类型可能有数十种，行业分类可能有数百种。在这种情况下，整理频数分布时可以根据需要，将某些类型合并为一组，如将行业合并为三大产业。也有些时候，为了分析的需要，常常将定性数据整理为两组，如将行业分为高污染行业和低污染行业。

软件操作指南 2-1

COUNTIF 函数

在 Excel 中，可以借助 COUNTIF 函数来完成定性数据频数分布的构建。该函数的语句为 COUNTIF(数值区间，条件)。例如，若性别变量的数据位于 B2:B31，则统计男性的频数时，写入函数“COUNTIF(B2:B31,1)”，就可输出男性的数量。

二、定量数据频数分布的构建

对于定量数据，虽然也可以视每个取值为一组，但由于其取值通常较多，所以一般采取将若干取值合并为一组的方法，即先将数据的取值范围划分为若干个区间，然后确定各个区间出现的观察值的数目，最后编制出类似定性数据的频数分布表。

在构建定量数据的频数分布时，可以按照以下几个步骤进行。

（1）确定原始数据的极差。所谓极差，是指最大值和最小值之间的差。

（2）确定组数。分组的一个基本原则是保持组数为 5～15，因为组数太少或太多，得到的频数分布对数据的归纳都不是最优的。分组的组数有一个 Sturges 经验公式：$m=1+3.322\lg n$，其中 m 为组数，n 为数据个数。当然，最终的组数是由研究者决定的。研究者会根据极差，选择一个既能覆盖全部数据，又比较有实用价值的组数。

（3）确定组距。用极差除以组数，就得到组距。一般，人们会将得到的组距进行四舍五入，使之变为整数。此外，人们还习惯于比较整齐的组距，如 5 和 10。

（4）确定分组。以一个等于或小于最小值的数值作为第一组的起点（即第一组的下组限），以第一组的下组限加一个组距得到的数值作为第一组的终点（即第一组的上组限）；以第一组的上组限作为第二组的下组限，以第二组的下组限加一个组距得到的数值作为第二组的上组限；依此类推，得出各组的下组限和上组限，且最后一组以一个等于或大于最大值的数值作为上组限，这样就完成了一个频数分布。通常，第一组的下组限和最后一组的上组限都是组距的整数倍，这样便于计算各组的上下组限。

需要注意的是，由于相邻两组的组限有重合，所以在分组时应保证一个数值必须属于且只能属于一个组。通常采用的是“上组限不在内”的原则，即各组的上组限不属于本组。例如，如果上一组为 10～20，下一组为 20～30，则 20 不属于上一组，而属于下一组。

定量数据的频数分布中还有另外三个基本概念，即组中值、频率和累积频数（或累积频率）。

（1）组中值。组中值是指代表每组数值中间水平的值，它等于各组上下组限的平均数，其计算公式为：

$$组中值=\frac{下组限+上组限}{2} \tag{2-1}$$

（2）频率。频率是指某一给定组的频数占总频数的比例，其计算公式为：

$$频率=\frac{组频数}{总频数} \tag{2-2}$$

（3）累积频数。累积频数是指频数分布中小于等于某一组上组限的数值的频数总和，其计算公式为：

$$第\ k\ 组的累积频数=\sum_{i=1}^{k}第\ i\ 组的频数 \tag{2-3}$$

由式(2-3)可以得出，最后一组的累积频数等于总频数。用同样的计算方法可以得到累积频率，最后一组的累积频率等于 1。

【例 2-2】 某电力企业 2006—2010 年的月度电费收入数据如表 2-6 所示。

表 2-6 某电力企业 2006—2010 年的月度电费收入

单位：亿元

年份 / 电费 / 月份	2006 年	2007 年	2008 年	2009 年	2010 年
1 月	3.3	3.6	3.3	3.9	8.9
2 月	2.9	3.5	3.1	4.0	8.1
3 月	4.6	3.6	3.9	3.9	8.8
4 月	4.2	5.0	5.9	5.2	8.9
5 月	7.1	6.7	7.1	8.7	17.0
6 月	7.8	10.7	8.7	6.3	26.2
7 月	8.7	10.8	10.9	10.6	27.6
8 月	6.4	11.4	11.3	11.3	26.7
9 月	7.0	10.9	10.9	10.6	26.9
10 月	7.8	9.3	7.3	15.3	17.8
11 月	5.2	5.9	8.1	12.5	17.4
12 月	4.2	3.5	4.4	10.5	10.0

要求：构建上述数据的频数分布，并计算组中值、频率和累积频率。

解 （1）确定极差。极差＝27.6－2.9＝24.7 亿元。

（2）确定组数。根据 Sturges 经验公式可得，$m=1+3.322\lg 60\approx 7$，因此组数取 7。

（3）确定组距。组距＝24.7÷7≈4 亿元。

（4）确定分组。第一组的起始点必须是 2.9 亿元或者更小，这样才能将数据中的最小值包含在内；最后一组的终点必须是 27.6 亿元或者更大，这样才能将数据中的最大值包含在内。本例中，频数分布以 0 为起点，以 28 亿元为终点。最终得到的频数分布如表 2-7 所示。

表 2-7　电费收入的频数分布

组/亿元	频　数	组中值/亿元	频　率	累积频率
0～4	11	2	0.18	0.18
4～8	19	6	0.32	0.50
8～12	21	10	0.35	0.85
12～16	2	14	0.03	0.88
16～20	3	18	0.05	0.93
20～24	0	22	0	0.93
24～28	4	26	0.07	1.00
合计	60	—	1.00	—

表 2-7 的频数分布揭示了该电力企业电费收入的大体分布状况，其中大部分月份的电费收入在 4 亿～12 亿元，尤其是 8 亿～12 亿元出现的频率最高。

实际中构建的定量数据频数分布会出现以下几种特殊的情形。

第一，开口组。如果第一组表述为“××以下”，或者最后一组表述为“××以上”，或者两者同时出现，则这种组属于开口组。由于无法对开口组计算组中值，所以给一些统计量的计算带来困难。建议在构建频数分布时，应尽量避免设置开口组。

第二，不等距分组。如果不同组的组距不相等，则这种分组称为不等距分组。不等距分组常常出现在某个或某些组的频数过低的情形。例如，有以下 20 个数据。

1　1　1　2　2　2　3　3　3　11　18　18　19　19　19　19　19　20　20　20

若按照惯例，则形成的频数分布如表 2-8 所示。

表 2-8　组距为 4 的等距分组

组	频　数
0～4	9
4～8	0
8～12	1
12～16	0
16～20	7
20～24	3
合计	20

通过表 2-8 可以看到，在这个频数分布中，第二组、第三组和第四组的频数非常小，其中第二组和第四组的频数为零。如果数据量很大，且数据在某些区间上的分布很稀疏，则空白组或稀疏组的数目就会较多。对于这种情形，可以将多个相邻组合并，形成组距不等的结果。例如，表 2-8 的频数分布经过合并后变为表 2-9。

表 2-9 不等距分组

组	频 数
0～4	9
4～16	1
16～20	7
20～24	3
合计	20

不等距分组会影响直方图的绘制，关于这一点，下文会有详述。

需要注意的是，尽管定量数据的频数分布可以依据一定的准则进行构建，但对于相同的原始数据，不同的研究者构建的频数分布可能有所不同，甚至相差很大。例如，对于上面的例子，有的研究者可能会构建组距为 10 的频数分布（见表 2-10）。

表 2-10 组距为 10 的等距分组

组	频 数
0～10	9
10～20	8
20～30	3
合计	20

从某种意义上讲，频数分布的构建会因人而异，这可以认为是统计应用具有艺术性的一个表现。

软件操作指南 2-2

FREQUENCY 函数

在 Excel 中，可以借助 FREQUENCY 函数完成定量数据频数分布的构建。该函数的语句为 FREQUENCY(数值区间，分组区间)。其中，分组区间只需要输入各个区间的上组限即可。注意 FREQUENCY 函数的输出结果是一个数组，因此要先选中一个输出区域，而且写入函数后，需要同时按下 Ctrl、Shift 和 Enter 键，才能在选中的输出区域中输出整个频数分布。例如，对于例 2-2，假如数值区间位于电子表格的 B2:F13，在 H2:H8 区域输入各组上组限，选中 I2:I8，写入函数“=FREQUENCY(B2:F13,H2:H8)”，同时按下 Ctrl、Shift 和 Enter 键，则在电子表格的 I2:I8 中会出现各组频数。

使用 Excel 的 FREQUENCY 函数时，一定要注意 Excel 软件对于相邻两组组限的确定采取了“上组限不在内”的原则，因此在输入分组区间时必须格外谨慎。例如，在例 2-2 中，需要输入的上组限应为 3.99，7.99，…，27.99，而不能是 4，8，…，28。

第二节　统计数据的图示

一般来说，人们认识事物通常是由表及里进行的，对统计数据的认识也不例外。将统计数据以统计图的形式展现出来，通过统计图直观地认识统计数据的一些基本特征，分析者可以得到很多启发，从而为进一步的统计分析奠定基础。因此，统计数据的图示方法是统计应用者的必备技术。统计图的主要任务是对频数分布进行展示，由于不同类型的统计数据具有不同的特点，所以其图示方法也有所不同。

一、定性数据的图示

与定量数据相比，定性数据最大的特点在于：第一，定性数据表现为文字，确切地说是类型，因此其在数轴上仅仅能够表现出类型的差异，而不能以精确的坐标示之；第二，定性数据的文字属性决定了其取值通常只有有限多个，而且这些取值可以区分开来，因此可以将这些取值在数轴上分别列示。

与定性数据上述两个特点相适应的统计图是条形图、柱形图和饼图，此外，对于定序数据，还可绘制累积频数分布图。

（一）条形图与柱形图

条形图的绘制方法是：以组（即类）为纵轴，每个组对应一个水平的条形；以频数（或频率）为横轴，每个条形的长度对应于各组的频数（或频率）。柱形图的绘制方法与条形图相似，只需要将横轴和纵轴交换。

【例 2-3】　对例 2-1 的性别数据绘制条形图。

解　例 2-1 的性别条形图如图 2-1 所示。

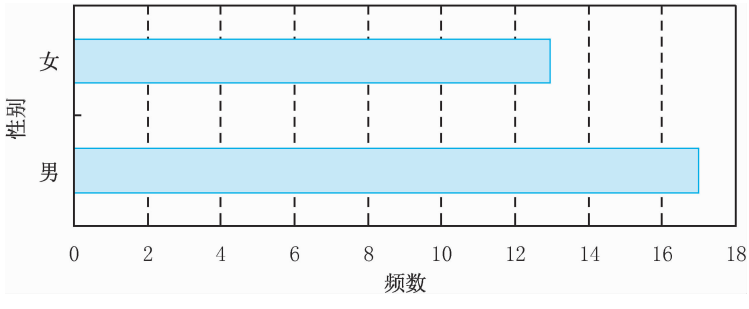

图 2-1　性别条形图

由图 2-1 可以清楚地看到被调查者的性别分布，男生明显多于女生。

【例 2-4】　对例 2-1 的满意度数据绘制柱形图。

解　例 2-1 的满意度柱形图如图 2-2 所示。

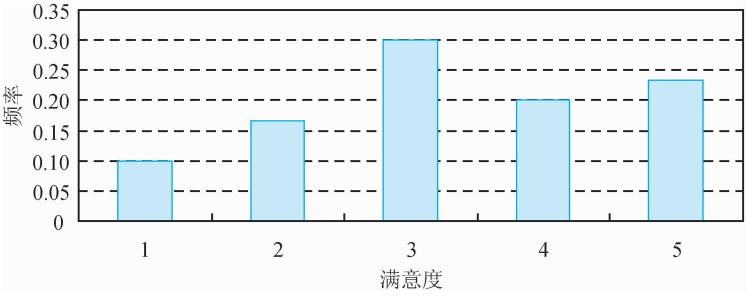

图 2-2 满意度柱形图

由图 2-2 可以看到，对图书馆的满意度一般的学生最多，非常满意的学生人数仅次于满意度一般的学生。总的来看，满意度高的学生要多于满意度低的学生。

（二）饼图

饼图的绘制方法为：绘制一个圆形饼，将饼分割为若干扇形对应于各组，扇形面积的比例对应于各组的频数（或频率）。

【例 2-5】 对例 2-1 的满意度数据绘制饼图。

解 例 2-1 的满意度饼图如图 2-3 所示。

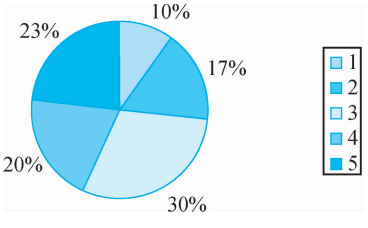

图 2-3 满意度饼图

（三）累积频数分布图

由于定序数据可以计算累积频数，所以对于定序数据，还可以绘制累积频数分布图。累积频数分布图的绘制方法为：以组为横轴，以累积频数为纵轴，将组与累积频数在坐标平面上的交点连接起来，得到一条折线，即为累积频数分布图。

【例 2-6】 对例 2-1 的满意度数据绘制累积频数分布图。

解 例 2-1 的满意度累积频数分布图如图 2-4 所示。

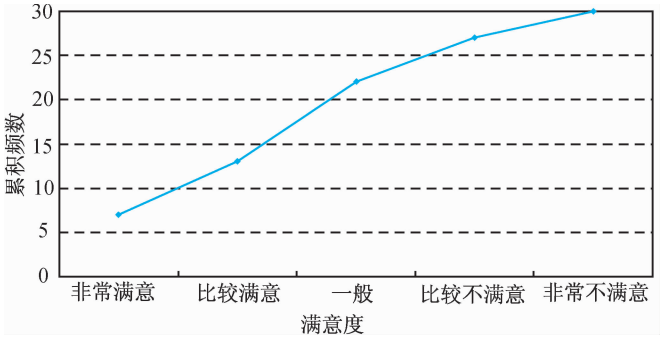

图 2-4 满意度累积频数分布图

二、定量数据的图示

与定性数据相比，定量数据的特点为：第一，定量数据表现为数字，因此其在数轴上能够以精确的坐标示之；第二，定量数据通常有多个取值，因此需要先将这些取值进行分组，再在数轴上加以显示。

定量数据的图示方法主要是直方图和茎叶图。

（一）直方图

等距分组数据直方图的绘制方法为：在二维坐标平面上绘制一些柱形，每组对应一个柱形，柱形在横轴上的具体位置由各组的上下组限所决定，柱形的高度对应于各组的频数（或频率）密度。

所谓频数（或频率）密度，是指用各组的频数（或频率）除以各组组距所得到的商。根据直方图的绘制方法，易知直方图中柱形的宽度表示各组组距，而各个柱形的面积等于各组的频数（或频率）。采用频数（或频率）密度可以消除组距不同的影响。如果纵轴采用频率，则所有柱形的面积之和等于总频数；如果纵轴采用频率，则所有柱形的面积之和等于 1。

对于等距分组的数据而言，由于各组组距相同，因此绘制直方图时常常直接以频数（或频率）作为纵轴，此时柱形面积正比于各组频数（或频率）。

【例 2-7】 根据表 2-7，对例 2-2 的电费收入数据绘制直方图。

解 表 2-7 是等距分组，因此可以以频数为纵轴，从而得到电费收入直方图如图 2-5 所示。

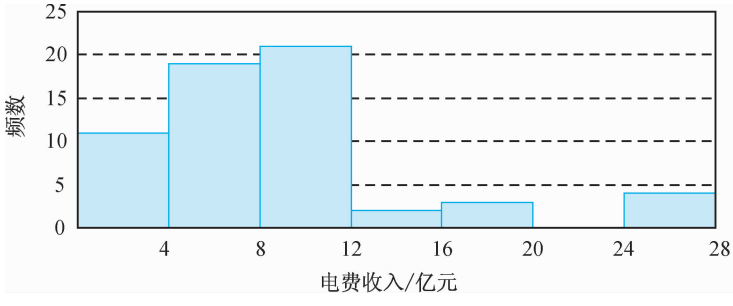

图 2-5　电费收入直方图一

由图 2-5 可以看到，电费收入多数集中在 12 亿元以下。

需要注意的是，直方图和柱形图外观比较相似，但是两者有着本质上的差异。具体表现为以下几点。

（1）柱形图的横轴只表示类的差异，而没有具体的坐标。例如，图 2-2 中的五个柱形可以平移到横轴的任何位置，对图所传达的信息并没有任何影响，该图中横轴的数字只是类的标志而已。直方图的横轴则有真正的坐标，柱形的位置不可以随意移动。由此导致柱形图中柱形的宽度没有意义，而直方图中柱形的宽度则代表每一组的组距。

（2）柱形图中柱形之间既可以分离，也可以相接，其排列方式由绘图者决定；直方图中柱形之间有无间隔则不是随意决定的，而是由每一组的区间位置决定的。例如，图 2-2 中的五个柱形完全可以相接起来；而图 2-5 中的前 5 个柱形必须相接，第 5 个柱形与第 6 个柱形

则必须分隔。

(3) 柱形图的纵轴表示频数;而直方图的纵轴只有在等距分组的情形下才表示频数,在非等距分组的情形下表示的则是频数密度。换言之,柱形图是通过柱形高度的对比反映各组频数的大小,而直方图则是通过柱形面积的对比来反映各组频数的大小。

【例 2-8】 对表 2-9 中的数据绘制直方图。

解 表 2-9 中的数据为不等距分组,因此在绘制直方图时纵轴应采用频数密度,如图 2-6(a)所示。

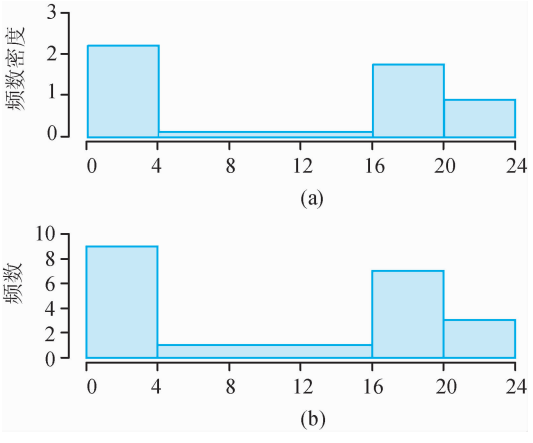

图 2-6 表 2-9 中的数据直方图

图 2-6(b)给出了以频数表示的直方图。显然,直观来看,图 2-6(b)会给人一个错误的印象,即误以为 4～16 组的频数为 3,原因在于以频数为纵轴夸大了柱形的面积。相比之下,图 2-6(a)才正确地反映了数据的分布。

(二) 茎叶图

茎叶图是直方图的"近亲",与直方图一样,茎叶图也能够概括数据的分布特征。两者最大的差别在于茎叶图是利用排序后的原始数据绘制的,而直方图是利用分组数据绘制的。

茎叶图的绘制包括以下几个步骤。

(1) 将原始数据排序。

(2) 选择适当的相邻的两个数位,如百位和十位,将每个数值在这两个数位之间分开。例如,123 可以分为 1|23。

(3) 绘制一条竖线,将每个数的高数位数值放置在竖线的左侧,称为"茎";将低数位数值放置在竖线的右侧,称为"叶"。注意每个数值的"叶"应当长在其"茎"所在的行。

绘制茎叶图的关键在于第二步。数位的选择取决于茎叶图行数的选择,实质上也就是组数或组距的选择。与构建频数分布时组数的确定具有主观性类似,茎叶图中行数的选择也依赖于绘图者的主观判断。一个经验准则为:$L=[10\times\lg n]$。其中,L 为行数,n 为数据个数,$[x]$表示不超过 x 的最大整数。例如,如果 $n=20$,则 $L=[10\times\lg 20]=[10\times 1.3]=13$。行数确定之后,需要确定每一行的数值区间。最简单的办法是用数据的极差除以行数,再以与商最接近的 10 的幂作为区间长度。例如,如果极差为 40,行数为 13,则 $40\div 13\approx 3.08$,与 3.08 最接近的 10 的幂(即 10^0)为 1,因此取区间长度为 1。根据区间长度便可确定区分茎叶

的数位。例如，如果区间长度为 1，则茎取个位，叶取十分位；如果区间长度为 100，则茎取百位，叶取十位。

此外，如果横行的叶子太多，图形过于拥挤，还可以将横行分裂，使每个茎重复。

【例 2-9】　对例 2-2 的电费收入数据绘制茎叶图。

解　由于数据个数为 60，因此行数 $L=[10\times\lg 60]=[10\times 1.78]=17$。电费收入的极差为 24.7，$24.7\div 17\approx 1.45$，与 1.45 最接近的 10 的幂为 1，因此取区间长度为 1，进而确定取个位数为茎，十分位数为叶。

最终绘制的茎叶图如图 2-7 所示。

茎/亿元	叶/亿元										数值个数
2	9										(1)
3	1	3	3	5	5	6	6	9	9	9	(10)
4	0	2	2	4	6						(5)
5	0	2	2	9	9						(5)
6	3	4	7								(3)
7	0	1	1	3	8	8					(6)
8	1	1	7	7	7	8	9	9			(8)
9	3										(1)
10	0	5	6	6	7	8	9	9	9		(9)
11	3	3	4								(3)
12	5										(1)
13											(0)
14											(0)
15	3										(1)
16											(0)
17	0	4	8								(3)
18											(0)
19											(0)
20											(0)
21											(0)
22											(0)
23											(0)
24											(0)
25											(0)
26	2	7	9								(3)
27	6										(1)

图 2-7　电费收入茎叶图

由图 2-7 可以看出，电费收入大多数集中在 2 亿～12 亿元之间，其分布有三个峰值，即 3 亿～4 亿元、8 亿～9 亿元和 10 亿～11 亿元，电费收入出现在这三个区间的频数最多。

容易发现，茎叶图很像倒置的直方图。与直方图相比，茎叶图最大的优点是保留了原始数据的信息，这样更有利于计算各种统计量。然而当数据量很大时，就不便于用茎叶图来表示了。

三、正确使用统计图

统计图的绘制具有很大的灵活性，但这也给误用和滥用统计图留下了很大的空间。下面提出几点正确使用统计图的建议。

（1）不要刻意改变图形形状，如调整图形的长宽比例或改变坐标轴刻度，以达到分析者自己预设的目标。

（2）绘制直方图时，尽量使用等距分组，否则容易落入以频数为纵轴的陷阱，造成对数据分布形态的错误判断。

（3）分组组距的选择非常重要，一定要慎重。以例 2-2 的电费收入直方图（见图 2-8）来进行说明。

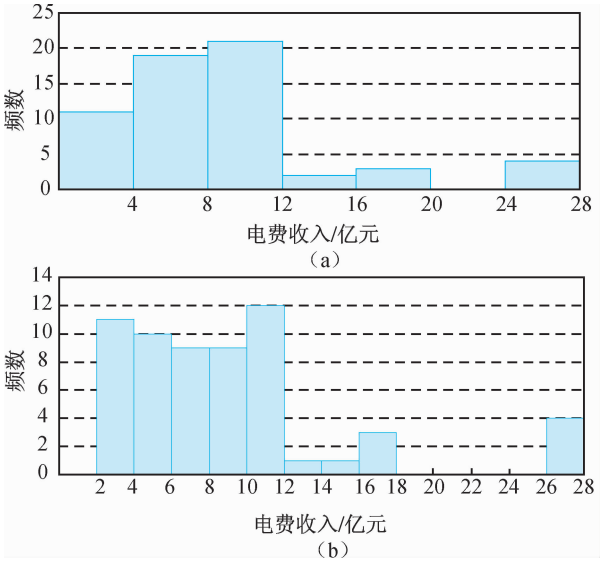

图 2-8　电费收入直方图二

图 2-8 中，两个直方图的差别在于组距不同，显然，两个图所揭示的数据结构也是不同的。其中，图 2-8(a)显示数据只在 8 亿～12 亿元出现一个峰值，而图 2-8(b)则显示数据在 2 亿～4 亿元和 10 亿～12 亿元有两个峰值。图 2-8(a)显示数据在 16 亿～20 亿元的分布有一定的频数，但图 2-8(b)则表明这一区间的数据只集中在 16 亿～18 亿元，事实上，电费收入从未出现在 18 亿～20 亿元。相比而言，组距较小的图 2-8(b)保留了更多的细节信息。

然而，组距也不是越小越好。不同的研究有不同的目标函数，通过使得目标函数最优化，可以确定最优组距。例如，使用直方图来估计数据的密度函数是直方图的重要应用之一，对此问题感兴趣的读者可以参考非参数统计的有关理论。对于初学者而言，需要尝试不同的组距，从中选择一个最能反映数据结构的直方图。

第三节　统计数据分布特征的测量

借助于频数分布和统计图，分析者可以大体上了解数据的分布形态和范围。但有时人们还希望从数据中提炼出几个关键的特征来概括数据的分布特征。常用的分布特征包括数据的集中趋势、离散程度、偏度系数以及峰度系数。

一、集中趋势的测量

所谓集中趋势，是指数据向数据中心靠拢的趋势。从集中趋势的测量中，可以了解一组数据的代表性水平。测量集中趋势的统计量包括众数（mode）、中位数（median）和均值（mean）。

（一）众数

所谓众数，是指一批数据中出现频数最多的数。显然，众数的计算不涉及数学运算，因此它适用于任何尺度的数据。

对于定量数据，如果只有分组数据，则需要对众数进行近似的估计，估计方法是以众数组的组中值作为分组数据的众数。所谓众数组，是指频数最大的组。

【例 2-10】　对于例 2-1 的数据，计算性别和满意度的众数。

解　由于男生的人数多于女生，所以性别的众数为男生；由于满意度一般的学生人数最多，所以满意度的众数为一般。

【例 2-11】　对于表 2-6 的数据，计算电费收入的众数。

解　由于 8 亿～12 亿元组的频数最多，是众数组，所以估计的众数为（8＋12）÷2＝10 亿元。

由于众数仅与数据的频数有关，所以不易受极端值的影响，统计学将这种性质称为稳健性（robustness）。例如，有 10 个学生的统计学考试成绩为：

95　95　95　95　95　95　95　90　90　90

则众数为 95。显然，这个成绩最适合代表这批成绩。而如果成绩为：

95　95　95　95　95　95　95　90　90　40

虽然出现了一个特别低的成绩 40，这是一个极端值，但是众数不受影响，依然是 95。

对于一批数据，众数可能是唯一的，也可能有多个，还可能不存在。以例 2-2 的电费收入数据为例，3.9 亿元、8.7 亿元和 10.9 亿元这三个数值出现的次数最多，都是三次，即这批数据有三个众数。而对于下面的数据：

1　2　3　4　5　6　7　8　9　10

则不存在众数。

如果一批数据有两个众数，称为双众数；如果有三个及三个以上的众数，称为多众数。

在实际经济社会活动中，研究者常常根据众数来作决定。例如，制衣厂往往只生产几种典型型号的衬衣、外套、套装等，而选举活动中则直接根据众数来确定当选者。

（二）中位数

所谓中位数，是指对一批数据排序之后，处于中间位置的数值。显然，在一批数据中，一半的数据比中位数大，另一半的数据比中位数小。

由于中位数的计算涉及排序，所以中位数适用于定序数据和定量数据，而不适用于定类数据。

【例 2-12】 对例 2-1 的数据，计算满意度的中位数。

解 由于中位数的位置为(30+1)÷2=15.5，位于一般组，所以满意度的中位数为一般。

对于定量数据，如果分析者掌握了原始数据，则首先确定中位数的位置，然后按照下面的公式计算：

$$Me=\begin{cases} X_{\frac{n+1}{2}} & \text{如果 } n \text{ 为奇数} \\ \dfrac{X_{\frac{n}{2}}+X_{\frac{n}{2}+1}}{2} & \text{如果 } n \text{ 为偶数} \end{cases} \tag{2-4}$$

式中，Me 代表中位数；$X_{\frac{n+1}{2}}$ 和 $X_{\frac{n}{2}}$ 表示原始数据排序后处于其下标数字所示位置上的数值。例如，按照由小到大的顺序排序，X_1 即最小值，X_n 即最大值。

【例 2-13】 对例 2-2 的电费收入数据计算中位数。

解 由于 n 为 60，所以中位数等于排序后第 30 个数值(7.8)与第 31 个数值(8.1)的平均数，即：

$$Me=(7.8+8.1)\div 2=7.95 \text{ 亿元}$$

如果只有分组数据，则可以用下面的公式近似地计算中位数：

$$Me=L_m+\frac{\frac{n}{2}-f_L}{f_m}\times i \tag{2-5}$$

式中，Me 代表中位数；n 为总频数；L_m 为中位数所在组的下组限；i 为中位数所在组的组距；f_m 为中位数所在组的频数；f_L 为中位数所在组的前一组的累积频数。

【例 2-14】 对表 2-7 的分组数据计算中位数。

解 中位数所在的位置为 $n/2=30$，在 4 亿～8 亿元组，$L_m=4$ 亿元，$i=4$，$f_m=19$，$f_L=11$。因此：

$$Me=4+\frac{30-11}{19}\times 4=8 \text{ 亿元}$$

与众数一样，中位数也具有稳健性。例如，将电费收入数据的最大值 27.6 替换为 276，中位数不变。而且，一批数据的中位数具有唯一性。此外，中位数还具有一个优良性质，即：

$$\underset{U}{\operatorname{argmin}}\sum_{i=1}^{n}|x_i-U|=Me \tag{2-6}$$

式中，x_i-U 度量了数值 x_i 相对于某个中心 U 的差异，统计学中称之为离差。式(2-6)的含义为：在所有可能的中心 U 中，能够使得所有数值的绝对离差之和最小的数据中心是中位数。这是一个非常优良的性质，在工程学中有广泛的应用。

（三）均值

均值也称为算术平均数，是指一批数据中所有数值之和除以总频数后的数值。由于均值的计算涉及加法，所以只能用于定量数据，而不适用于定性数据。

对于原始数据，均值的计算公式为：

$$\bar{x} = \frac{\sum_{i=1}^{n} x_i}{n} \tag{2-7}$$

式中，$\bar{x}$ 代表均值；x_i 代表各个数值；n 为数据个数。

【例 2-15】　对例 2-2 的电费收入数据计算均值。

解

$$\bar{x}=\frac{3.3+2.9+4.6+\cdots+17.8+17.4+10.0}{60}=9.1\text{ 亿元}$$

对于分组数据，需要借助组中值来计算均值，计算公式为：

$$\bar{x} = \frac{\sum_{i=1}^{K} x_i f_i}{\sum_{i=1}^{K} f_i} = \sum_{i=1}^{K} x_i \left(\frac{f_i}{\sum_{i=1}^{K} f_i} \right) \tag{2-8}$$

式中，$\bar{x}$ 代表均值；x_i 为各组的组中值；K 为组数；f_i 为各组的频数；$\frac{f_i}{\sum_{i=1}^{K} f_i}$ 为频率，又称为权数。

【例 2-16】　计算表 2-7 中电费收入的均值。

解

$$\bar{x}=\frac{2\times11+6\times19+\cdots+26\times4}{11+19+\cdots+4}=8.87\text{ 亿元}$$

与众数和中位数相比，均值对信息的利用最为充分，因此它是研究者在测量定量数据的集中趋势时最常用的统计量。从统计思想来看，均值是数值的误差相互抵消之后所呈现出来的事物的内在规律。容易推出均值具有以下数学性质：

$$\sum_{i=1}^{n} (x_i - \bar{x}) = 0 \tag{2-9}$$

式(2-9)表明，数据观察值相对于均值的误差是完全可以消除的。

但是，也正是由于均值的计算需要用到所有数值的信息，所以如果数据中存在极端值，就会对均值造成较大的影响，从而削弱均值对数据的代表性，即均值不具有稳健性。

例如，在前面的 10 个学生统计学考试成绩的例子中，如果成绩为：

95　95　95　95　95　95　95　90　90　90

则均值为 93.5。而如果成绩为：

95　95　95　95　95　95　95　90　90　40

则由于出现了一个极端值 40，使得均值明显降低，只有 88.5。显然，直观上看，93.5比 88.5 更能代表这批数据。

由此可知，在存在极端值的情况下，应当选择众数或中位数来测量定量数据的集中趋势；或者至少应当在计算均值的同时，提供众数或中位数的信息。

此外，均值还有一个性质，即：

$$\underset{U}{\operatorname{argmin}} \sum_{i=1}^{n} (x_i - U)^2 = \bar{x} \tag{2-10}$$

式(2-10)的含义为:在所有可能的中心 U 中,能够使得所有数值的离差平方和最小的数据中心是均值。这是一个非常优良的性质,在统计学中有广泛的应用。

专栏 2-1

均值的几种变形

(1) 几何平均数。在实际分析中,研究者常常会关注事物或现象的增长态势。例如,给定一定时期内的增长速度数据 $g_1, g_2, \cdots g_i, \cdots, g_t$,其中 $g_i = \frac{a_i - a_{i-1}}{a_{i-1}} = \frac{a_i}{a_{i-1}} - 1 (a_i$ 表示第 i 期某事物的发展水平)。如果采用算术平均数,则平均增长速度为:

$$\bar{g}_A = \frac{\sum_{i=1}^{t} g_i}{t}$$

这种计算方法似乎没有问题,但其计算出的 $\bar{g}_A$ 却并不能保证下式成立:

$$a_t' = a_0(1+\bar{g}_A)^t = a_t$$

上式是平均增长速度的内在要求。例如,若 $a_0 = 2, a_1 = 4, a_2 = 6$,则 $g_1 = 1, g_2 = 0.5, \bar{g}_A = 0.75$,而 $a_2' = 2 \times (1+0.75)^2 = 6.125 \neq 6$。此时,可以考虑利用几何平均数来计算平均增长速度。具体计算公式为:

$$\bar{g}_G = \sqrt[t]{\prod_{i=1}^{t}(1+g_i)} - 1$$

对于上面的例子,$\bar{g}_G = \sqrt{(1+1)\times(1+0.5)} - 1 = 0.73$,而 $a_2' = 2 \times (1+0.73)^2 = 6$。经过简单的推导,可得:

$$1 + \bar{g}_G = \exp\left\{\frac{\sum_{i=1}^{t}\ln(1+g_i)}{t}\right\}$$

先对原始数据进行对数变换,求出对数变换数据的算术平均数之后再还原为指数,即可得到几何平均数。因此,几何平均数实质上是算术平均数的一种变形。

(2) 截尾均值。由于均值不具有稳健性,所以如果数据中有极端值,可以考虑去除这些极端值,从而消除其对于均值的影响。去掉数据大小两端的若干数值后计算得出的均值称为截尾均值。截尾均值的计算公式为:

$$\bar{x}_t = \frac{x_{n\alpha+1} + x_{n\alpha+2} + \cdots + x_{n-n\alpha}}{n - 2n\alpha}$$

式中,$x_{n\alpha+1}, x_{n\alpha+2}, \cdots, x_{n-n\alpha}$ 为原始数据排序后处于其下标标示位置上的数值;α $(0 \leqslant \alpha < 0.5)$ 为截尾系数;n 为数据个数。如果 $\alpha = 0$,截尾均值就是普通的均值。容易看出,截尾均值实际上是对被截掉的数值赋予零的权数,因此它本质上也是均值的一种变形。

由于截尾均值既保留了均值对数据信息利用充分的优点,又解决了均值不稳健的缺点,因此其已被广泛用于评比竞赛的综合打分阶段。

例如，5 名裁判给某位选手的打分如下：

8　10　8　8　8

如果计算均值，则受一个高分 10 分的影响，该选手的平均分为 8.4；如果采用 $\alpha=0.2$ 的截尾均值进行计算，则 $\bar{x}_t=\frac{x_{(2)}+x_{(3)}+x_{(4)}}{5-2\times5\times0.2}=8$ 分，没有受到 10 分的影响。

二、离散程度的测量

数据分布特征的第二个方面是数据的分散程度，具体来说，是数据相对于集中趋势的离散程度。测量离散程度的统计量与测量集中趋势的统计量是相对应的，包括与众数相应的异众比率(variation ratio)、与中位数相应的四分位差(quartile deviation)以及与均值相应的方差(variance)和标准差(standard deviation)。将离散程度的测量值和集中趋势的测量值结合起来使用，可以更完整地描述数据的分布特征。

(一) 异众比率

所谓异众比率，是指非众数组的频数占总频数的比重。异众比率越大，表明数据相对于众数的离散程度越大，也意味着众数对数据的代表性越差。

【例 2-17】　计算例 2-1 中性别和满意度的异众比率。

解　由于女生的频数为 13，所以性别的异众比率为：

$$13\div30=43\%$$

由于一般组以外的频数共 21，所以满意度的异众比率为：

$$21\div30=70\%$$

相比而言，满意度的离散程度更大，可见用一般来代表满意度，其代表性较差。

(二) 四分位差

所谓四分位差，是指数据的上四分位数(upper-quartile)与下四分位数(lower-quartile)的差，也称为内距(inter-quartile range)。四分位数是指将数据频数平分为四份的三个数值，最大的四分位数称为上四分位数，中间的四分位数称为中位数，最小的四分位数称为下四分位数。显然，上、下四分位数之间有 50%的数据。

上、下四分位数的计算方法与中位数类似。在原始数据中，上、下四分位数分别对应由小到大排序后第$\frac{3(n+1)}{4}$和第$\frac{n+1}{4}$位置上的数值；如果位置不是整数，需要对相邻两个数值进行插值计算。

对于分组数据，其上、下四分位数需要进行近似计算。具体计算公式为：

$$Q_U=L_{Q_U}+\frac{\frac{3n}{4}-f_L}{f_{Q_U}}\times i \tag{2-11}$$

$$Q_L=L_{Q_L}+\frac{\frac{n}{4}-f_L}{f_{Q_L}}\times i \tag{2-12}$$

式(2-11)和式(2-12)中的符号含义与式(2-5)的符号类似，读者应能自行判断，此处不再

赘述。

【例 2-18】 计算表 2-7 中分组数据的四分位差。

解 上、下四分位数分别为：

$$Q_U = 8 + \frac{\frac{180}{4} - 30}{21} \times 4 = 10.86$$

$$Q_L = 4 + \frac{\frac{60}{4} - 11}{19} \times 4 = 4.84$$

因此，电费收入的四分位差为：10.86－4.84＝6.02。

四分位差反映了中间 50％数据的差异大小。如果四分位差很大，则数据围绕中位数的离散程度很大，表明中位数对数据的代表性较差。

专栏 2-2

箱 线 图

除了直方图和茎叶图之外，定量数据的频数分布还可以通过箱线图(box plot)进行描绘。箱线图是利用数据的最小值、下四分位数、中位数、上四分位数和最大值五个特征值绘制而成的，因其由一个箱子和两条线段组成而得名。

箱线图的绘制方法为：首先确定上述五个特征值，然后连接上、下四分位数绘出箱子，最后将最小值和最大值与箱子以线段相连接。箱线图的一般形式如图 2-9 所示：

图 2-9 箱线图的一般形式

在一个图中可以同时绘制多批数据的箱线图，从而方便数据的比较，这是箱线图的一大特点。图 2-10 就是一个多批数据的箱线图，可以看到，上面一批数据的中位数比下面一批数据高，即上面一批数据的平均水平高，但是它的四分位差也较大，因此其离散程度较大。

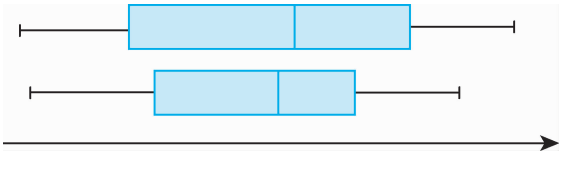

图 2-10 多批数据的箱线图

（三）方差和标准差

方差是指每个数值与均值的离差平方的平均数。对于原始数据，其计算公式为：

$$s_n^2 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n} \tag{2-13}$$

方差的另外一个计算公式为：

$$s_{n-1}^2 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n-1} \tag{2-14}$$

对于分组数据，方差的计算公式为：

$$s_n^2 = \frac{\sum_{i=1}^{K}(x_i - \bar{x})^2 f_i}{n} \tag{2-15}$$

$$s_{n-1}^2 = \frac{\sum_{i=1}^{K}(x_i - \bar{x})^2 f_i}{n-1} \tag{2-16}$$

式(2-15)和式(2-16)中，x_i 为各组的组中值；f_i 为各组的频数；K 表示分为 K 组。

【例 2-19】　计算例 2-2 中电费收入的方差。

解　根据式(2-13)和式(2-14)可得：

$$s_n^2 = \frac{(3.3-9.1)^2+(2.9-9.1)^2+\cdots+(10-9.1)^2}{60} = 35.47$$

$$s_{n-1}^2 = \frac{(3.3-9.1)^2+(2.9-9.1)^2+\cdots+(10-9.1)^2}{60-1} = 36.07$$

其中，s_n^2 与 s_{n-1}^2 的差别在于分母不同。在实际应用中，计算总体数据的方差时通常使用 s_n^2，并将总体方差记为 σ^2；计算样本数据的方差时则通常使用 s_{n-1}^2。可以证明，s_{n-1}^2 是总体方差 s_n^2 的无偏估计，而 s_n^2 则是有偏估计。当样本容量 n 很大时，两者的计算结果差别很小。

方差的正的平方根称为标准差。与 s_n^2 对应的标准差记为 s_n，与 s_{n-1}^2 对应的标准差记为 s_{n-1}。标准差区别于方差的一个主要特点为：标准差的单位和原始数据的单位一样，而方差的单位是原始数据单位的平方。

显然，方差或标准差越大，则数据围绕均值分散的程度越大，均值的代表性就越差。

虽然标准差可以单独使用，但是在实际中研究者经常将它与均值结合起来应用。常见的结合有两种：一是对数据进行标准化，二是计算离散系数。

标准化是指用数值减去均值的差再除以标准差，标准化值称为 Z 值。具体计算公式为：

$$Z_i = \frac{x_i - \bar{x}}{s} \tag{2-17}$$

【例 2-20】　计算例 2-2 中 2006 年 1 月和 2010 年 7 月电费收入的标准化值。

解　由于 $\bar{x}=9.1$，$s_{n-1}=\sqrt{36.07}=6.01$，因此可得，2006 年 1 月的电费收入 3.3 亿元的标准化值为 $\frac{3.3-9.1}{6.01}=-0.97$，而 2010 年 7 月的电费收入 27.6 亿元的标准化值为 $\frac{27.6-9.1}{6.01}=3.08$。

Z 值代表一个数值高于或低于该组数据均值的标准差的倍数，用 Z 值可以将数值距离

均值的原始距离转化为距离均值多少个标准差。例如，−0.97 表示数值低于均值 0.97 个标准差，而 3.08 则表示数值高于均值 3.08 个标准差。

根据经验法则，当数据大致为正态分布时，大约 68%的数据分布在距离均值一个标准差的范围内，大约 95%的数据分布在距离均值两个标准差的范围内，大约 99.7%的数据分布在距离均值三个标准差的范围内。

对不同数据的离散程度进行比较时，由于方差和标准差是以均值为中心计算出来的，如果不同数据的均值不同，则直接比较方差或标准差是不可取的，需要剔除均值不可比的因素之后再进行比较。离散系数就是为此而设计的，它是以标准差除以均值得出的，即：

$$CV=\frac{s}{\bar{x}} \tag{2-18}$$

【例 2-21】 计算例 2-2 中电费收入的离散系数。

解 由于 $\bar{x}=9.1$，$s_{n-1}=6.01$，所以 $CV=\frac{6.01}{9.1}=0.66$。

专栏 2-3

切比雪夫不等式

根据经验法则判断数据的分布区间，要求数据大体上呈正态分布。然而，当数据不服从正态分布或分布形状未知时，研究者应如何判断数据的分布区间呢？切比雪夫不等式适用于所有分布，而不用考虑分布的形状。因此，当数据不服从正态分布或分布形状未知时，可以使用切比雪夫不等式。其表达式如下。

令均值为 μ，方差为 σ^2，则有

$$P(|X-\mu|\geqslant k)\leqslant\frac{\sigma^2}{k^2}$$

切比雪夫不等式意味着：无论分布的形状如何，至少会有 $\left(1-\frac{1}{k^2}\right)$ 的数值落在距离均值 k 个标准差的范围内。根据切比雪夫理论，无论分布形状如何，至少有 75%的数据落在距离均值两个标准差的范围内，至少有 89%的数据落在距离均值三个标准差的范围内。

三、偏度系数和峰度系数的测量

（一）偏度系数

所谓偏度系数，是指测量数据分布对称性的统计量。具体计算公式为：

$$SK=\frac{\sum_{i=1}^{n}(x_i-\bar{x})^3}{ns^3} \tag{2-19}$$

通过式(2-19)可以看出，偏度系数是基于数值相对于均值的离差的三次方计算出来的。

如果数据的分布是完全对称的，那么正负离差的三次方就可以完全抵消，从而使得 $SK=0$，图形如图 2-11(a)所示。如果数据中有少数较大的值，则正负离差的三次方抵消不掉，且正值大于负值，从而 $SK>0$，称这种数据分布为正偏或右偏，图形如图 2-11(b)所示。同理，如果数据中有少数较小的值，则 $SK<0$，称这种数据分布为负偏或左偏，图形如图 2-11(c)所示。

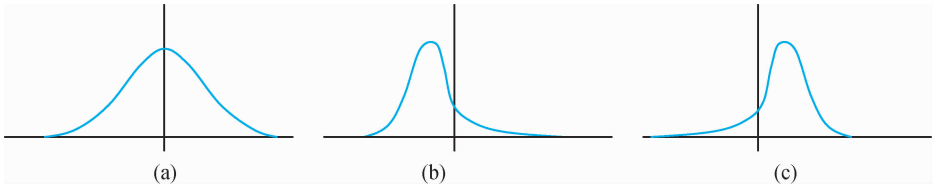

图 2-11　数据分布的对称性

例如，根据例 2-2 的电费收入数据可计算出 $SK=1.75$，表明电费收入为右偏分布。观察数据，确实可以发现有部分较大的电费收入数据。

（二）峰度系数

所谓峰度系数，是指测量数据分布相对于正态分布的扁平程度的统计量。其计算公式为：

$$K=\frac{\sum_{i=1}^{n}(x_i-\bar{x})^4}{ns^4}-3 \tag{2-20}$$

由式(2-20)可以看出，峰度系数是基于数值相对于均值的离差的四次方计算出来的。如果有少数数值远离均值，则 K 值会非常大。正态分布的 $K=0$。如果 $K>0$，称这种数据分布为尖峰分布；反之，如果 $K<0$，称这种数据分布为平峰分布。例如，根据例 2-2 的电费收入数据可计算出 $K=0.04$，表明电费收入为尖峰分布，意味着有极端值存在，符合对数据的观察。

软件操作指南 2-3

Excel 中的描述统计工具

在 Excel 中运用描述统计，有以下两种方法。

(1) 使用相应的统计函数。相应的统计函数包括：众数(mode)、中位数(median)、四分位数(quartile)、均值(average)、总体方差(varp)、样本方差(var)、总体标准差(stdevp)、样本标准差(stdev)、偏度系数(skew)、峰度系数(kurt)。

(2) 使用分析工具中的描述统计。首先打开 Excel 工作表，执行“工具”→“加载宏”→“分析工具库”；然后在“工具”下拉菜单中选择“数据分析”，从工具列表中选择“描述统计”，可直接输出常用的各种统计量。

专栏 2-4

卡尔·皮尔逊与偏斜分布①

卡尔·皮尔逊(Karl Pearson,1857—1936 年)是现代统计学的奠基人之一。在皮尔逊之前,科学所处理的事情都是真实的。开普勒试图发现行星如何在空间运行的数学规律,威廉·哈维的实验打算确定血液如何在某一特定动物的静脉和动脉中流动。然而,开普勒所试图追踪的“行星”实际上是一组数据,用来给地球上的观测者所看到的天空中微弱的光点定位;单匹马身上血液通过静脉流动的实际情形也许与在另一匹马或者一个人身上所可能看到的不同。这些研究处理的实际上都是观测到的数据,与数学测绘上的理想世界不一样,现实世界所获得的测量数据通常是真实值的近似值。皮尔逊指出,这些观测到的现象是随机的,是不真实的,所谓的真实是概率分布。科学中真实的东西并不是人们所能观测到或者把握到的,它们只是通过用来描述人们所观测事物随机性的数学函数来反映。

皮尔逊发现了被他称为偏斜分布(skew distribution)的一组分布函数,他宣称这组函数可以描述科学家在数据中可能遇到的任何分布类型。这组函数中的每一个分布都由四个参数(均值、标准差、偏度和峰度)所确定,这四个参数才是人们在科学研究中真正想确定的。从某种意义上讲,人们永远不能确定这四个参数的真实数值,只可能从资料中估计它们。

尽管后来的研究者发现,皮尔逊的分析系统具有一定的局限性,在许多情况下它并不适用。但皮尔逊对于分布形态的分析方法是现代统计学得以发展的基础。他关于分布函数和参数的思想统治了 20 世纪的科学,并在 21 世纪仍然保持着优势。

引例解析

对于本章开头提到的东北地区降水量特征问题,表 2-1 中的各个统计量提供了丰富的信息。具体包括以下几个方面。

(1) 从降水量的集中趋势来看。根据均值可以看到,随着纬度升高,年平均降水量呈减少趋势。

(2) 从降水量的绝对离散程度来看。根据离散系数的定义,可以计算出哈尔滨、长春和沈阳降水量的标准差分别为 119.8 mm、126.8 mm 和 149.1 mm;根据最大值和最小值计算极差,则哈尔滨、长春和沈阳降水量的极差分别为 698.2 mm、640.8 mm 和723.8 mm。可以看到,三个地区降水量的绝对离散程度存在一定差异,其中沈阳降水量的离散程度最大。

(3) 从降水量的相对离散程度来看。哈尔滨、长春和沈阳降水量的离散系数非常接近,

① 萨尔斯伯格.女士品茶[M].邱东,等,译.北京:中国统计出版社,2004.

亨德森哈里.数学——描述自然与社会的有力模式[M].王正科,等,译.上海:上海科学技术文献出版社,2008.

表明三个地区降水量的相对离散程度无明显差异。

(4) 从降水量的分布形态来看。由峰度系数可知，哈尔滨和长春的降水量是尖峰分布，而沈阳的降水量是平峰分布；由偏度系数可知，三个地区的降水量都为右偏分布。

根据以上特征，研究者可以对哈尔滨、长春和沈阳三个地区的降水量作进一步的分析，并加以预测，从而为区域经济发展的决策提供支持。

主要术语

频数分布　频数　组中值　频率　累积频数　极差　组距　下组限　上组限
条形图　柱形图　饼图　累积频数分布图　直方图　茎叶图　箱线图　集中趋势
众数　中位数　均值　离散程度　异众比率　四分位差　方差　标准差　标准化
离散系数　偏度系数　峰度系数

思 考 题

(1) 什么是频数分布？
(2) 定性数据频数分布的构建与定量数据频数分布的构建有何差别？
(3) 定性数据与定量数据的图示方法各有哪几种？
(4) 柱形图与直方图的区别是什么？
(5) 茎叶图有什么优缺点？
(6) 找一个应用统计图的例子，判断其分析过程有无错误。
(7) 众数、中位数和均值各自的优缺点是什么？
(8) 为什么要对数据进行标准化？
(9) 如何绘制和使用箱线图？
(10) 如果数据呈右偏或左偏分布，众数、中位数和均值的大小关系如何？

练 习 题

(1) 某班有 40 名学生，其中有 5 名担任学生干部。现在要推选一名优秀学生干部，每名学生可以提名 2 个学生干部，选票情况如表 2-11 所示。

表 2-11　学生选票情况

学生编号	提名 1	提名 2	学生编号	提名 1	提名 2
1	班长	学习委员	21	团支部书记	班长
2	班长	团支部书记	22	班长	生活委员
3	学习委员	团支部书记	23	学习委员	团支部书记
4	班长	体育委员	24	班长	生活委员
5	班长	生活委员	25	班长	学习委员
6	班长	学习委员	26	班长	团支部书记

续表

学生编号	提名 1	提名 2	学生编号	提名 1	提名 2
7	学习委员	生活委员	27	生活委员	班长
8	团支部书记	班长	28	班长	学习委员
9	团支部书记	班长	29	班长	体育委员
10	班长	生活委员	30	班长	学习委员
11	班长	团支部书记	31	班长	团支部书记
12	生活委员	班长	32	生活委员	团支部书记
13	体育委员	团支部书记	33	体育委员	学习委员
14	班长	团支部书记	34	班长	体育委员
15	班长	学习委员	35	班长	生活委员
16	生活委员	体育委员	36	生活委员	团支部书记
17	团支部书记	班长	37	团支部书记	学习委员
18	生活委员	学习委员	38	团支部书记	班长
19	学习委员	班长	39	班长	学习委员
20	班长	学习委员	40	班长	体育委员

根据上述数据，构建优秀学生干部提名情况的频数分布，据此绘制柱形图和饼图，并说明是依据哪个统计量来确定优秀学生干部的人选的。

（2）某研究小组通过电话调查收集了某地区家庭的常住人口数据。假设随机抽取 30 个家庭，每个家庭的人数如下：

3 3 1 2 5 2 4 1 3 4 2 3 1 2 3

2 3 1 2 3 4 2 3 2 6 3 2 1 3 3

根据上述数据，计算众数、中位数、均值、极差、内距和标准差。

（3）国家外汇管理局公布的 2009 年 12 月人民币对美元的汇率如表 2-12 所示。

表 2-12　2009 年 12 月人民币对美元的汇率

日　期	汇　率	日　期	汇　率	日　期	汇　率
2009－12－31	682.82	2009－12－21	682.83	2009－12－9	682.79
2009－12－30	682.83	2009－12－18	682.84	2009－12－8	682.77
2009－12－29	682.82	2009－12－17	682.81	2009－12－7	682.78
2009－12－28	682.82	2009－12－16	682.8	2009－12－4	682.72
2009－12－25	682.83	2009－12－15	682.78	2009－12－3	682.7
2009－12－24	682.85	2009－12－14	682.79	2009－12－2	682.68
2009－12－23	682.87	2009－12－11	682.77	2009－12－1	682.7
2009－12－22	682.85	2009－12－10	682.76		

注：以上数据以 100 美元为基数。

根据上述数据，计算众数、中位数、均值、极差、内距和标准差，并绘制直方图、茎叶图和箱线图。

（4）根据《2005 年度中国对外直接投资统计公报》，我国各行业 2005 年对外直接投资流量数据如表 2-13 所示。

表 2-13　我国各行业 2005 年对外直接投资流量数据

单位：万美元

行　业	投资额	行　业	投资额
农、林、牧、渔业	10 536	房地产业	11 563
采矿业	167 522	租赁和商务服务业	494 159
制造业	228 040	科学研究、技术服务和地质勘查业	12 942
电力、煤气及水的生产和供应业	766	水利、环境和公共设施管理业	13
建筑业	8 186	居民服务和其他服务业	6 279
交通运输、仓储和邮政业	57 679	卫生、社会保障和社会福利业	0
信息传输、计算机服务和软件业	1 479	文化、体育和娱乐业	12
批发和零售业	226 012	公共管理和社会组织	173
住宿和餐饮业	758		

根据以上数据回答下列问题。

① 计算投资额的均值、方差和标准差。

② 分别计算制造业和建筑业的 Z 值，并解释各个 Z 值的含义。

③ 计算偏度系数和峰度系数，讨论数据的分布形状。

（5）2006 年我国主要城市年平均相对湿度的数据如表 2-14 所示。

表 2-14　2006 年我国主要城市年平均相对湿度

单位：%

城　市	年平均	城　市	年平均	城　市	年平均
北京	53	合肥	72	贵阳	79
天津	61	福州	72	昆明	69
石家庄	55	南昌	71	拉萨	35
太原	58	济南	58	西安	67
呼和浩特	47	郑州	62	兰州	56
沈阳	68	武汉	71	西宁	56
长春	59	长沙	72	银川	51
哈尔滨	57	广州	71	乌鲁木齐	54
上海	70	南宁	76		
南京	71	海口	78		
杭州	71	重庆	75		

利用以上数据回答下列问题。

① 计算均值、中位数和众数，指出三个统计量中哪一个更适合用来反映这批数据的集中趋势，并说明原因。

② 计算异众比率、内距、方差和标准差。

③ 北京市的 Z 值是多少？贵阳市的 Z 值是多少？解释这些 Z 值的含义。

④ 计算偏度系数和峰度系数。

⑤ 绘制箱线图。

(6) 假设某市场研究员调查了 170 个怀旧歌曲的听众，获得的数据如表 2-15 所示。

表 2-15 怀旧歌曲听众分布表

年龄/岁	频 数
15～20	9
20～25	16
25～30	27
30～35	44
35～40	42
40～45	23
45～50	7
50～55	2

利用以上数据回答下列问题。

① 怀旧歌曲听众年龄的均值、中位数和众数是多少？

② 怀旧歌曲听众年龄的方差和标准差是多少？

③ 怀旧歌曲听众年龄的分布是否对称？

④ 估计一下在怀旧歌曲听众总体中，95%的听众的年龄位于哪个区间？

⑤ 绘制怀旧歌曲听众年龄的直方图。

(7) 某上市公司股票 2009 年 11 月和 12 月的收盘价如表 2-16 所示。

表 2-16 某上市公司股票 2009 年 11 月和 12 月的收盘价

日 期	收盘价/(元/每股)	日 期	收盘价/(元/每股)
2009－12－31	6.19	2009－11－30	5.93
2009－12－30	6.09	2009－11－27	5.85
2009－12－29	5.92	2009－11－26	5.92
2009－12－28	5.87	2009－11－25	6.13
2009－12－25	5.82	2009－11－24	6.1
2009－12－24	5.85	2009－11－23	6.28
2009－12－23	5.76	2009－11－20	6.26
2009－12－22	5.74	2009－11－19	6.34
2009－12－21	5.84	2009－11－18	6.34
2009－12－18	5.86	2009－11－17	6.25

续表

日　　期	收盘价/(元/每股)	日　　期	收盘价/(元/每股)
2009－12－17	5.96	2009－11－16	6.14
2009－12－16	6.03	2009－11－13	5.99
2009－12－14	6.14	2009－11－12	5.95
2009－12－11	6.03	2009－11－11	5.98
2009－12－10	6.04	2009－11－10	6.02
2009－12－9	6	2009－11－9	5.98
2009－12－8	6.08	2009－11－6	5.99
2009－12－7	6.2	2009－11－5	6
2009－12－4	6.2	2009－11－4	5.96
2009－12－3	5.97	2009－11－3	5.96
2009－12－2	6.01	2009－11－2	5.96
2009－12－1	5.96		

请分别计算该公司股票 11 月和 12 月收盘价的中位数、均值、方差、标准差、离散系数、偏度系数和峰度系数，并对两个月的股票价格进行简要的比较分析。