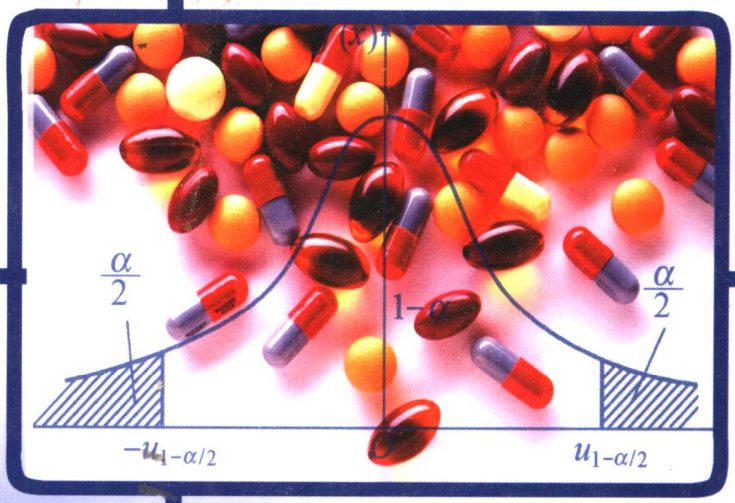
21 世纪高等医药院校教材

医药应用概率统计

高祖新　韩可勤　主编



科学出版社
www.sciencep.com

21世纪高等医药院校教材

医药应用概率统计

主　编　高祖新　韩可勤
编　者　高祖新　韩可勤　尹　勤
　　盛海林　易登录　王　菲

科学出版社
北京

内 容 简 介

本书是编者在所进行的“统计及应用课程教改研究和实践”研究成果荣获国家级教学成果二等奖基础上，所编写的颇具特色的概率统计基础课程教材。

全书针对医药本科学生的基础和培养要求，适当选取教材内容的深度和广度，内容系统全面，例题典型实用，编写力求简明易懂，深入浅出，富有启发性，应用性强。其内容包括数据处理、概率论基础、数理统计原理、知识及常用统计方法等，并在每章的最后给出 Excel 软件对应统计功能的操作应用，辅之适当的思考与练习题、应用统计软件的上机训练题和习题及参考答案。同时每章配有内容提要与综合举例，对各章核心内容以简表形式高度概括，并精选综合性典型例题和考研真题进行详解和分析，使本教材能兼顾本科基础课教学和更高教学要求（如考研）不同层次的需求。

本书主要适合于医药类各专业概率统计、数理统计、应用统计等基础课教材或参考书，也可作为各类非理工科专业同类课程的参考书，还可作为其他教学，如考研复习辅导及数学建模等的参考书。同时也是医药卫生工作人员颇为实用的统计应用参考书。

图书在版编目（CIP）数据

医药应用概率统计/高祖新，韩可勤主编. —北京：科学出版社，2005.8
（21 世纪高等医药院校教材）
ISBN 7-03-016112-2

Ⅰ. 医… Ⅱ. ①高… ②韩… Ⅲ. 数理统计-应用-医药学-医学院校-教材 Ⅳ. R311

中国版本图书馆 CIP 数据核字（2005）第 090738 号

责任编辑：胡治国 吴茵杰/责任校对：钟 洋
责任印制：刘士平/封面设计：黄 超

科学出版社 出版
北京东黄城根北街16号
邮政编码：100717
http://www.sciencep.com
双青印刷厂 印刷

科学出版社发行 各地新华书店经销

*

2005 年 8 月第 一 版 开本：787×1092 1/16
2005 年 8 月第一次印刷 印张：25
印数：1—5 000 字数：575 000

定价：39.00 元

（如有印装质量问题，我社负责调换〈环伟〉）

前　言

随着我国医药高等教育改革和发展的不断深入，高校医药人才的培养模式和要求有了极大的变化，同时也对我国医药类相关专业的统计及应用课程的教学提出了更高更新的要求。为此，我们进行了“统计课程结构、教学手段及面向专业的教学改革”重点教改课题的专题研究，对国内外概率统计及应用课程的教材和内容等进行深入的分析研究，同时结合学科发展动态、医药领域统计应用的特点要求和多年教改实践经验，对课程结构、内容、教学模式和教材等各方面进行了有益的改革探索，所取得的教改成果“统计及应用课程教改研究和实践”于2001年荣获国家级教学成果二等奖。在此基础上我们编写了这本颇具特色的《医药应用概率统计》课程教材。

概率统计是研究随机现象统计规律性的学科。本教材的编写既考虑到概率统计学科知识结构的科学性和系统性，又结合了医药领域对统计应用的具体要求和特点，同时针对医药本科学生的基础和培养要求，适当选取教材内容的深度和广度，并反映学科发展的时代特征，内容系统而全面，例题典型而实用，编写力求简明易懂，深入浅出。其主要特点是：

一、作为医药类相关专业基础课教材，在尽量保持概率统计学科的科学性和系统性的前提下，以“夯实学科基础，掌握概念方法，强化专业应用，培养实用技能”为重点，不片面追求理论的推导和证明，而强调理论与实际的结合，体现了学以致用的目的，并充分考虑更高层次（如硕士研究生入学考试）的教学要求。

二、所选内容涵盖概率论基础、医药应用领域数据处理和统计分析的基本原理、基本知识和常用统计方法，在系统而简明地介绍概率论知识的基础上，以统计数据的处理和分析为核心，注重统计方法思想和实际医药应用的阐述，结合数据和医药专业应用实例说明统计方法的特点、应用条件和场合等，从而形成以概率论基础、统计原理、统计方法及统计软件应用为主体并面向医药领域实际应用的内容体系。

三、强化以计算机应用为基础的统计技能的培养。现代医药领域数据处理和统计分析离不开计算机统计软件的应用，根据医药院校本科学生所用软件应满足普及、实用的要求，本教材选用了最为常用的 Microsoft Office 系统的 Excel 软件统计模块来进行统计软件应用的教学，操作指导具体详实，便于自学，并配有上机训练题，从而能真正提高读者运用统计工具分析和解决实际问题的实际操作能力，达到学以致用的目的。

四、目前已有日趋增多的医药相关专业（诸如医药经济类、医药管理类和医药工程类等专业）研究生录取都需要参加全国“数学”硕士研究生入学考试，为适应这些更高层次的教学要求，本教材内容已完全包含参加全国“数学”硕士研究生入学考试教学大纲中关于概率统计知识要求的内容，并在每章配有内容提要与综合举例，对核心内容以简表形式进行高度概括，并精选综合性典型例题和考研真题进行详解和分析，为准备考研等更高学习要求的读者提供极为有效的、兼具基础性和技巧性的解题指导。

五、由于本书内容系统全面，阐述深入浅出，用例经典实用，概括高度精炼，并兼顾本科基础课教学和更高教学要求不同层次的需求，使之不仅适用于医药类各相关专业概率统计、数理统计、应用统计等基础课教材或参考书，而且还可用于农林经管等各类非理工科专业同类课程的参考书，并可作为其他教学，如考研辅导及数学建模等的参考书。而书中极具实用性的医药统计的 Excel 软件操作应用，使之成为从事医药研究和工作的相关人员不可多得的统计应用参考书。

教材的主要内容有统计数据的描述和概括、随机事件和概率、随机变量及其分布、随机向量及其分布、大数定律与中心极限定理、抽样分布、参数估计、假设检验、方差分析、相关分析与回归分析、正交设计与均匀设计等章节，并在每章的最后一节给出 Excel 软件对应概率统计功能的操作应用，同时辅之以相应的典型实例、适当的思考与练习题和应用统计软件的上机训练题，以帮助学生消化、巩固所学内容，真正掌握统计应用的原理和方法。而在每章最后配有的内容提要与综合举例，不仅为读者重点掌握各章的核心内容提供便利，同时也为有更高要求的读者在综合性和技巧性解题方面的提高给予有效的帮助指导，使本教材能适应不同层次教学要求的需要。

本教材供一学期教学使用，也可根据课时和教学要求的不同选用部分内容。其中每章的最后一节——Excel 软件的统计应用，则可根据课时、计算机设备等客观条件来灵活取舍或安排自学。

本教材共分 11 章，其中第 1 ~ 6 章和各章 Excel 软件应用、内容提要等部分由高祖新负责编写，第 7 ~ 11 章由韩可勤负责编写，最后共同统稿纂定。

本教材的编著，得到了有关专家的关心和帮助，并参考了大量的教材和文献，在此表示衷心的感谢。由于时间和水平有限，书中疏漏和不妥之处在所难免，恳请各位读者批评指正。

编　者

2005 年 7 月

目　　录

第一章

统计数据的描述和概括

概率论和数理统计是从数量侧面来研究随机现象统计规律性的学科。由于随机现象的普遍性，使得概率论和数理统计在工农业生产、社会经济和现代科学技术各领域中具有极其广泛的应用，而这些应用同时也推动着概率论和数理统计这门学科不断发展和完善。

统计作为一种社会实践活动由来已久，其含义也较丰富。它可以指统计数据的搜集活动，即统计工作；也可以指统计活动的结果，即统计数据；还可以指分析统计数据的方法和技术，即统计学。**统计学(Statistics)**是关于研究对象的数据资料进行搜集、整理、分析和解释，以显示其总体特征和统计规律性的科学。在英文中，“Statistics”以单数名词出现时表示统计学，而以复数名词出现时则表示统计数据或统计资料。可见，统计学与统计数据是密不可分的，它是一门有关统计数据的科学，其主要方面包括数据搜集也就是取得统计数据，是进行统计分析的基础；数据整理则是用图表等形式来展示数据特征，使数据更加系统化、条理化，从而便于统计分析；数据分析是利用描述统计和推断统计等统计方法来研究数据，是统计学的核心；而数据解释则是对统计分析结果进行说明和应用。

目前，统计学已发展成为由若干分支学科组成的学科体系，它作为一门方法论的科学，也已应用到自然界和人类社会的各个领域。而从统计方法的构成来看，统计学可分为描述统计和推断统计。**描述统计(Descriptive statistics)**是搜集、整理和描述数据资料的方法，即研究如何取得反映客观现象的数据资料，并通过统计图表、统计指标等有效形式对数据资料进行整理和概括显示，进而得出反映客观现象的规律性特征。如果在统计研究中可以得到研究对象的全体即整个总体，则应用描述统计学就足够了。但是，实际研究中常常只能得到总体的一部分(称为样本)，这就需要根据这些样本的有限的、不确定的信息，利用概率论的理论来对总体进行科学的推断。**推断统计(Inferential statistics)**就是研究如何利用样本数据资料来推断总体数量特征的方法，它是在对样本数据进行描述的基础上，对统计总体的未知数量特征作出以概率形式表述的推断。描述统计和推断统计是统计方法的两个组成部分，描述统计是统计学的基础，推断统计是现代统计学的主体和核心内容。

由于在自然科学和社会科学的研究领域中，都需要通过数据处理和分析来解决实际问题，因而统计应用几乎扩展到了所有的科学研究领域。应用到经济领域就形成了经济统计学；应用到医药领域就形成了医药统计学；等等。例如在药学领域，无论是新药研制、药物鉴定、药理分析、试验设计，还是药政管理、处方筛选、医药信息等各个方面，都需要进行大量的数据资料的整理和分析。因此，有关医药概率统计的知识和必要的统计技能训练，是每个医药科技工作者必不可少的专门知识和技能，其学习和掌握对于有效而正确地利用数据资料进行医药领域的研究和实践具有极为重要的意义。

第一节 统计数据的计量尺度和类型

一、统计数据的计量尺度

统计数据(Data)是对客观现象计量的结果,是我们利用统计方法进行分析的基础。按照对事物计量的精确程度,可将所采用的计量尺度从低级到高级分为四个层次(表 1-1):定类尺度、定序尺度、定距尺度和定比尺度。采用不同的计量尺度可以得到不同类型的统计数据,适用于不同的统计分析方法。

表 1-1 四种计量尺度数学特性的比较

计量尺度		定类尺度	定序尺度	定距尺度	定比尺度
数学特性	分类(=,≠)	√	√	√	√
	排序(<,>)		√	√	√
	间距(+,-)			√	√
	比值(+,-,×,÷)				√

(1) **定类尺度**(或**分类尺度 Nominal measurement**):按照事物的某种属性对其进行平行的分类或分组,是最低级的计量尺度。该尺度只测度了事物之间的类别差,其计量结果只是表现为某种类别,可以计算各类别中元素或个体的频数。例如:人口的性别(男、女),血型(O、A、B、AB 型)或药物的种类等。

(2) **定序尺度**(或**顺序尺度 Ordinal measurement**):是对事物之间等级或顺序差别的一种尺度。它不仅可以将事物分成不同的类别,还可以可以确定这些类别的优劣和顺序。其计量结果虽然也是表现为类别,但类别间可比较顺序,可以进行大小的比较,但不能进行加减乘除的数学运算。例如:新药的等级(一类、二类、……),考试的等级成绩(优、良、中、及格、不及格)等等。

(3) **定距尺度**(或**间距尺度 Interval measurement**):是对事物类别或次序之间间距的尺度,它不仅能将事物区分为不同类型并进行排序,还可以准确指出类别之间的差距。该尺度通常使用自然或物理单位作为计量尺度,其计量结果表现为数值,并可以计算差值,进行数学的加减运算。例如:对不同地区温度的测量(华氏和摄氏)等。

(4) **定比尺度**(或**比率尺度 Ratio measurement**):与定距尺度属于同一尺度层次,其计量结果也表现为数值。它除了具有前面三种尺度的所有特性外,还能够计算两个测度值之间的比值,并用绝对固定的"零点"表示不具备所测量的特性,即用"0"表示"没有"或"不存在",可以进行加减乘除的数学运算。例如:"百分制"的考试成绩;医药企业销售收入;人的身高、体重等等。注意:定距尺度与定比尺度差别在于不存在绝对零点,故只能比较数值差,而不能计算比值。现实生活中,我们主要使用定比尺度。

上述 4 种计量尺度对事物的测量层次是由低级到高级、由粗略到精确逐步递进的。高层次计量尺度具有低层次计量尺度的全部特性,能够计量低层次尺度所计量的事物,反之不成立。其差异见下表。

在统计分析中，应该尽量用高层次的计量尺度，因为计量尺度层次越高，所含的统计信息和数学特性就越多，所用统计方法也就越多，分析也就越方便。

二、统计数据的类型

（一）统计数据的分类

统计数据是我们采用计量尺度所得的计量结果。对应于上述四种计量尺度，统计数据就可以分为**定类数据（Nominal data 或分类数据 Categorical data）**、**定序数据（Ordinal data 或顺序数据 Rank data）**、**定距数据（Interval data）**、**定比数据（Ratio data）**等四种类型，通常又将其归为以下两大类数据：

（1）**定性数据（Qualitative data）**：又称**品质数据**，包括定类数据和定序数据，说明的是事物的品质特征，不能用数量表示，其结果通常表现为类别。

（2）**定量数据（Quantitative data）**：又称**数值型数据（Numerical data）**，包括定距数据和定比数据，说明的是事物的数量特征，是用数值来表示的，其结果通常表现为具体数字。

区分尺度的层次和数据的类型非常重要，如下表 1-2 所示，对不同类型的数据必须采用不同的统计方法来进行处理和分析。

表 1-2　不同数据类型的统计应用之比较

数据类型		定性数据（品质数据）		定量数据（数值型数据）	
		定类数据	定序数据	定距数据	定比数据
表现形式		类别（无序）	类别（有序）	数值（+ −）	数值（+ − × ÷）
应用统计量	众数	√	√	√	√
	中位数		√	√	√
	均值			√	√
	方差、标准差			√	√
应用统计方法		计算各组频数等，进行列联表分析等非参数法		计算各种统计量，进行参数估计、假设检验、回归分析、方差分析法等参数法	
对应变量		定类变量	定序变量	数值变量（离散变量、连续变量）	

实际问题中绝大多数数据资料是定量数据，本书所介绍的统计方法也主要用于定量数据的分析处理，只有非参数方法等可用于定性数据的研究。虽然只有定量数据可转化为定性数据，但也可以通过每类赋值（即编码）的方法，使定量数据的统计分析方法应用于定序数据。

（二）两类统计数据的转换

根据分析的需要，定量数据与定性数据之间经常要做转换。

（1）定量数据的定性化转换：例如，学习成绩由百分制转化成五等级制，这时定量数据就成了定性数据。

（2）定性数据的数量化转换：很多情况下，数据需要计算机处理。为了便于计算机的识别和运算，对定性数据可以赋值进行数量化转换。例如，性别是属于定性数据变量，可将男

女分别取值为1和2。取值1和2之间没有量的差别，只是一种“数据代码”。如果文化程度是按文盲、小学、初中、高中、大学及以上分组，此变量属于定序数据变量，可分别取值为0,1,2,3,4。此时取值0,1,2,3,4之间不仅是一种“数据代码”，也有量的区别。

三、变量及其类型

在统计中，将说明现象的某种属性或标志称为**变量**(**Variable**)，对变量进行测量或观察的值称为**观察值**(**Observation**)或**变量值**。统计数据就是统计变量的观察值。根据变量的记录形式分别为定类数据、定序数据和数值数据，相应地变量可以分为**定类变量**(**Nominal variable**)、**定序变量**(**Ordinal variable**)和**数值变量**(**Numerical variable**)。

数值变量中，如果变量可以取有限个值，并可以一一列举，称为**离散变量**(**Discrete variable**)，如制药公司数、仪器个数等。如果数值变量可以取无穷多个值，其取值是连续不断的，不能一一列举，就称为**连续变量**(**Continuous variable**)，如时间、温度、产品尺寸等。实际应用时，当离散变量的取值很多时，也可以当作连续变量来处理。

由于在实际中，应用最多的是数值变量，大多数统计方法所处理的也都是数值变量，故我们一般将数值变量简称为变量，即通常所说的变量主要是数值变量。

四、统计数据的搜集

搜集数据资料是统计分析的基础，而数据资料的来源有两个：

(1) 直接来源：通过专门组织的调查或科学试验来采集**原始数据资料**(**Primary data**)。其中专门调查是取得社会经济数据的重要手段，科学试验是取得自然科学数据的重要手段。

(2) 间接来源：利用已公开出版(报道)的信息资料或尚未公开的信息资料来搜集**次级资料**(**Secondary data**)，包括图书资料和报刊杂志、广播电视等媒体和互联网中的各种数据资料，使用时应注意数据的含义、计算口径和方法，并在引用时注明数据来源。

第二节 统计数据的整理和显示

统计数据的整理就是根据统计研究的任务，对搜集到的数据资料进行科学的加工和汇总，使数据资料系统化，以反映研究总体的特征、规律和趋势。数据整理和显示通常包括数据的审核筛选、分类或分组、汇总、列出统计图表等。

一、定性数据的整理和显示

在对统计数据进行整理时，首先要进行数据的审核筛选，以保证数据的质量。然后再根据不同的数据类型进行处理。

对于定性数据(品质数据)主要作分类整理。定性数据包括定类和定序数据，其数据本身就是对事物的一种分类和类别排序，只需按不同数据(类别)进行分组，算出各组的**频数**(**Frequency**)或**频率**(**Relative frequency**)，列出**频数分布表**(**Frequency table**)即可。例如，表1-3就是根据2000年我国人口普查数据得到的对我国6周岁以上人口按受教育程度分组形成的频数分布表。

表 1-3　2000 年我国 6 周岁以上各种受教育程度的人口数

受教育程度	文盲、半文盲	小学	初中	高中及中专	大专及以上
人数(亿)	1.1093	4.5191	4.2989	1.4109	0.4571
百分比(%)	9.4	38.3	36.4	12.0	3.9

数据来源:《中国人口统计年鉴 2001》,中国统计出版社,第 46 页

利用上表数据,我们就可得到 2000 年我国各种受教育程度的人口数的**垂直条形图**(**Vertical bar chart**),它直观地反映了我国各种受教育程度的人口分布状态(图 1-1)。

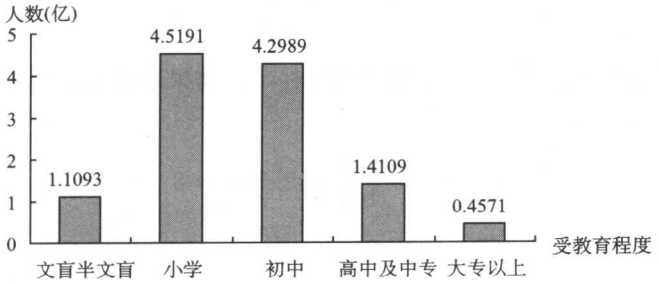


图 1-1　2000 年我国 6 周岁以上人口的各种受教育程度的垂直条形图

条形图(**Bar chart**)和**圆形图**(**Pie chart**)(见本章第四节)是反映定性数据或离散变量数据的分布特征和构成比的常用统计图形,在定性数据的统计图表显示中起着很好的作用。在本章第四节还列出了由表 1-3 的数据画出的水平条形图和圆形图(第四节中图 1-5,图 1-6)。

二、定量数据的整理和显示

定量数据资料统计整理的目的是了解定量数据的分布规律和类型,并根据分布类型选用适当的统计指标描述其集中趋势和离散趋势等统计特征。其整理和显示主要包括按数量标志进行分组,编制频数分布表,并采用**直方图**(**Histogram**)及**频数折线图**(**Frequency polygon**)等统计图形来表示其结果,以更直观清晰地表示其频数分布状态。

定量数据统计分组时,对于离散变量且变量值较少时,可按每个变量值为一组进行单变量值分组;对于连续变量或变量值较多时,应采用将全部变量值依次划分为若干个区间,每个区间为一组的组距分组。在组距分组中,一个组的最小值、最大值分别称为该组的下限、上限。

例 1.1　现有某高校某专业 110 名学生某门课的成绩(分)数据如下,试编制频数分布表来进行数据的整理显示。

76　42　94　97　72　88　55　96　62　83　99　80　81　77　68　90　67　85　69　61
76　73　81　65　61　87　87　93　88 100　89　99　65　61　74　97　62　72　91　49
72　82　98 100　73　51　71　99　68　94　82　85　79　74　55　87　49　85　72　78
97　86　53　71　73　90　88　77　80　86　71　96　85　46　73　66　98　55　98　81

79 84 86 74 86 62 74 79 59 96 97 69 89 86 81 78 84 99 45 95
82 91 67 73 89 89 84 74 32 72

解:以该例数据整理为例,给出定量数据组距分组法编制频数分布表的步骤。

(一) 确定组数

组数 k 的确定应以能够显示数据的分布特征和规律为目的,一般设 5 ~ 15 组,可根据数据本身的特征和数据的个数来定。通常当数据个数小于 50 时,可分为 5 ~ 6 组;当数据个数为 100 左右时,可分为 7 ~ 10 组;当数据个数超过 500 时,可分为 10 ~ 15 组。在实际分组时,也可按 Sturges 提出的经验公式来定组数 k:

$$k = 1 + \frac{\ln N}{\ln 2}$$

其中 ln 为以 e 为底的自然对数,N 为数据个数,对计算结果取整数后即是组数,在实用中可参考使用。在本例中,

$$k = 1 + \frac{\ln N}{\ln 2} = 7.781 \approx 8$$

即大致可分为 8 组。

(二) 确定组距

在分组中,组距 d 是指该组上限与下限之差,一般多采用等组距。此时,组距 d 可以由全部数据的最大值、最小值和组数 k 来定:

$$d = \frac{\text{最大值} - \text{最小值}}{\text{组数 } k} \text{(取整)}$$

取整是为了便于数据整理。本例中,最大值 = 100,最小值 = 32,故组距

$$d = \frac{100 - 32}{8} = 8.5$$

为便于计算,组距一般取 5 或 10 的倍数,而且第一组的下限应低于数据的最小值,最后一组的上限应该不低于数据的最大值。因此,本例中组距 d 取 10,首组下限为 30,实际分组数是 7 组。

(三) 分组计算频数,形成频数分布表

对上面数据进行分组,采用手工划记法或计算机汇总,计算各组频数,列出频数分布表,见表 1-4。

表 1-4 学生成绩数据频数分布表

成绩分组	30 ~ 40	40 ~ 50	50 ~ 60	60 ~ 70	70 ~ 80	80 ~ 90	90 ~ 100
频数 f	1	5	6	15	27	32	24
频率 f/n	0.009	0.045	0.0545	0.136	0.245	0.291	0.218

组距分组时,应该遵循“不重不漏”的原则。即数据在计入分组频数时,不重复不遗漏。对连续变量采用相邻两组组限重叠时,一般规定“组上限不在内”,只有最后一组包括上限。如在上表分组中,“30 ~ 40”表示[30,40),即上限 40 在分组时不计入该组,而应该计入下一

组。另外为避免出现空白组(数据频数为0)或个别极端值被漏掉,第一组和最后一组可以采用开口组"××以下"及"××以上",开口组通常以相邻组的组距作为其组距。

有时,为反映各组数据的一般水平,通常用**组中值(Middle point value)**作为该组数据的代表值,即

$$\text{组中值} = \frac{\text{下限值} + \text{上限值}}{2}$$

组中值在利用频数分布表数据进行均值、方差等计算或制作频数折线图时将起作用。

为了统计分析需要,有时需要观察某一数值以下(或以上)的频数或频率之和,即计算出**累积频数**或**累积频率(Cumulative frequency)**,如下列表1-5所示。

表1-5　学生成绩数据累积频数分布表

成绩分组	30~40	40~50	50~60	60~70	70~80	80~90	90~100
组中值	35	45	55	65	75	85	95
累计频数	1	6	12	27	54	86	110
累计频率	0.009	0.055	0.109	0.245	0.491	0.782	1.000

(四)结果的统计图示

为了显示定量数据的整理结果,一般用直方图和频数折线图等专门用于显示分组数据频数分布特征的统计图,以便直观全面地认识和分析定量数据的分布特征和规律。这里仅列出根据频数分布表1-4所绘制的直方图,相应的(累积)频数折线图等参见第四节。

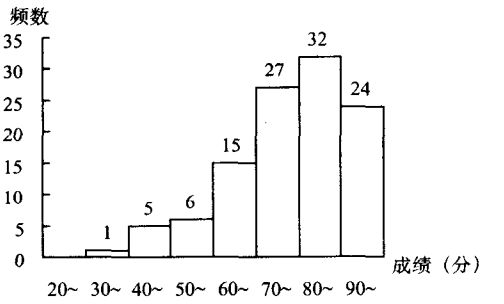


图1-2　学生成绩数据的频数直方图

三、常用统计软件的应用

在实际处理时,尤其是对于数据量较大的实际问题,一般通过计算机利用有关统计软件进行有关数据整理和统计图表显示等工作。目前常用的统计软件主要有SAS(统计分析系统)、SPSS(社会科学统计软件)、Excel(电子表格)等。

(一) SAS(统计分析系统)

SAS系统,全称Statistical Analysis System(统计分析系统),是模块化、集成化的应用软件系统,具有完备的数据管理、数据分析、数据存取、数据显示等功能,除统计分析外还有制图、矩阵运算、运筹规划、质量控制和医药临床研究等功能,为医药研究、经济管理、社会科学、自然科学等各领域的众多用户所采用,是当前最为流行的国际标准通用的分析统计软件,但其操作略为繁琐。

(二)SPSS(社会科学统计软件)

SPSS,全称Statistical Package for Social Science(社会科学统计软件),具有操作简便、统计功能并全、数据交换强大以及视窗组合等特点,也是当前最为流行的统计分析软件,在商

务、政府部门、教学与科研单位的定量研究中发挥了巨大的作用。

（三）Excel（电子表格软件）

Excel 是一个功能强大但使用简单的电子表格软件。Excel 在数据组织、数据管理、数据计算、数据分析及图表分析等方面提供了强大的功能；利用它的“函数”和“数据分析”功能也可以进行各种简单的和基本的统计运算和分析。

由于 Excel 软件普及程度高，操作运算也较为简便，本书主要介绍 Excel 软件的统计分析与运算处理的操作，以提高和拓展数据处理和统计分析的应用能力。

第三节　数据分布特征的统计概括

从数据的频数分布表或直方图等可以看到数据分布的两个重要特征：**集中趋势**（**Central tendency**）和**离散趋势**（**Disperse tendency**）。集中趋势是指数据向其某一中心值靠拢的倾向；离散趋势是指数据远离其中心值的趋势，也称为离中趋势。集中趋势和离散趋势反映了数据分布特征的两个重要侧面，各有其相应的统计代表值即测度值，又称统计量。

一、数据分布集中趋势的测度

描述数据分布集中趋势的测度值即统计量主要有均值、众数和中位数，又被称为数据分布的位置度量，其中最重要的是均值。

（一）均值

均值（**Mean**）也称为算术平均值，是全部数据的算术平均，记为 $\bar{x}$。均值是数据分布集中趋势的最主要测度值，在统计学中具有重要的地位。它适用于定量数据，不能用于定性数据。均值的计算公式将根据数据形式的不同而不同。

对未经分组整理的原始数据，设数据观察值为 x_1，x_2，…，x_n，均值的计算采用简单算术平均数公式：

$$\bar{x} = \frac{x_1 + x_2 + \cdots + x_n}{n} = \frac{1}{n}\sum_{i=1}^{n} x_i \tag{1.1}$$

例如，对例 1.1 中的原始数据，计算 110 名学生考试成绩的均值为

$$\bar{x} = \frac{76 + 42 + 94 + \cdots + 72}{110} = \frac{8591}{110} = 78.10$$

在 Excel 中，计算均值用其函数公式“ = AVERAGE(A1 : T6)”即可得结果（下页图 1-3）

对分组整理的数据，设原始数据被分为 k 组，各组的组中值为 m_1，m_2，…，m_k，各组观察值出现的频数分别为 f_1，f_2，…，f_k，其中 $\sum_{i=1}^{k} f_i = n$，均值的计算采用加权算术平均数公式：

$$\bar{x} \approx \frac{m_1 f_1 + m_2 f_2 + \cdots + m_k f_k}{f_1 + f_2 + \cdots + f_k} \approx \frac{1}{n}\sum_{i=1}^{k} m_i f_i \tag{1.2}$$

例 1.2　根据前面频数分布表 1-4 中的数据，试计算 110 名学生成绩的均值。

解：计算过程如下所示