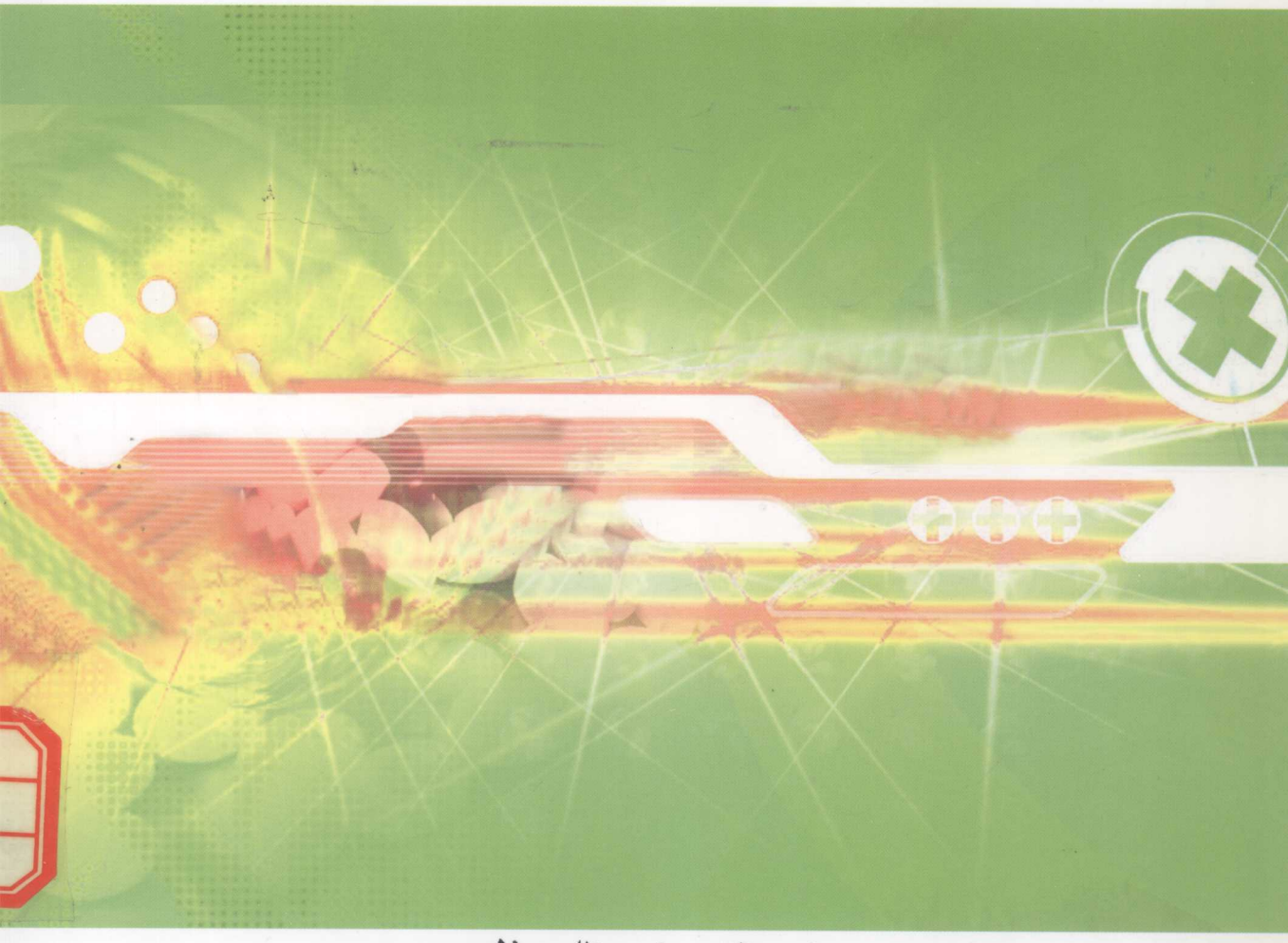
普通高等教育“十一五”国家级规划教材

全国高职高专药学类专业规划教材

医药应用统计

（第二版）

主编　高祖新　尹　勤



科学出版社
www.sciencep.com

普通高等教育“十一五”国家级规划教材
全国高职高专药学类专业规划教材

医药应用统计

（第二版）

主　编　高祖新　尹　勤
副主编　徐　伟　徐　宁
编　委　（按姓氏汉语拼音排序）
高祖新（中国药科大学）
麻佳蕾（金华职业技术学院）
王凤侠（枣庄科技职业学院）
徐　宁（山东药品食品职业学院）
徐　伟（沈阳药科大学高职学院）
尹　勤（南京人口管理干部学院）

科学出版社
北　京

内 容 简 介

本书是教育部普通高等教育"十一五"国家级规划教材和全国高职高专药学类专业规划教材之一,是在第一版的基础上修订而成。教学内容更加切合高职教学的实际,结构体系也更为合理完善。教材全面介绍了医药应用领域的数据处理与图表呈现;统计应用所必需的概率基础;数理统计的基本原理、基本概念和基本知识;常用统计推断和统计分析方法;用 Excel 进行数据处理与统计分析的实际操作应用等。主要包括数据的描述与统计概括、概率论基础、抽样分布、参数估计、假设检验、方差分析、相关与回归分析、正交试验设计八章。各章正文以医药应用案例贯穿全程,并附有学习目标、小结、目标检测、链接及目标检测参考答案等。全书还附有统计用表、课程教学基本要求和中英文名词索引等,并有配套的 PPT 教学课件,以方便教师教学,全面提升学生的学习能力。

本书可作为医药高职高专各专业学生学习医药统计(或数理统计)等基础课程的教材或教学参考书,也可供各类专业人员特别是医药卫生工作者学习参考。

图书在版编目(CIP)数据

医药应用统计 / 高祖新,尹勤主编. —2 版. —北京:科学出版社,2009

普通高等教育"十一五"国家级规划教材·全国高职高专药学类专业规划教材

ISBN 978-7-03-026288-2

Ⅰ. 医… Ⅱ. ①高…②尹… Ⅲ. 医用数学-数理统计-高等学校:技术学校-教材 Ⅳ. R311

中国版本图书馆 CIP 数据核字(2009)第 238251 号

策划编辑:邱　波 / 责任编辑:邱　波 / 责任校对:刘小梅

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
骏走印刷厂印刷
科学出版社发行　各地新华书店经销
*
2004 年 9 月第　一　版　　开本:787×1092　1/16
2009 年 12 月第　二　版　　印张:15
2009 年 12 月第九次印刷　　字数:356 000
印数:26 001—31 000
定价: 27.00 元
(如有印装质量问题,我社负责调换)

第二版编写说明

《医药应用统计》第一版自2004年8月出版发行以来，得到了全国医药高职高专院校广大师生的欢迎和肯定，至今已经印刷7次，并在2006年被评为教育部“普通高等教育‘十一五’国家级规划教材”。

为适应近年来我国医药高职高专教育快速发展的需要，结合多年的教学实践和教材编写经验以及国内外优秀统计学教材的成果，我们对第一版的内容体系进行了全面的修订和完善。本次再版的主要特色有以下几点。

一、本次再版在保持第一版特色和优势的基础上，本着“夯实数理统计基础，强化药学应用背景，增强统计软件训练，提升自主学习能力”的编写指导方针，在保持数理统计学科系统性前提下，进一步加强统计理论与医药实际的结合，突出医药高职高专学生的针对性和统计软件应用的实用性，更好地体现“学以致用”的教学目的。

二、按照医药高职学生的培养目标和要求，适当选取教材内容的深度和广度。教材内容涵盖统计应用所必需的概率论基础；医药应用领域的数据处理与图表呈现；数理统计的基本原理、基本概念和基本知识；常用统计推断和统计分析方法、统计软件（Excel数据分析模块）的实际操作应用等。再版的教学内容更加切合高职教学的实际，结构体系也更为合理完善。同时，我们还编制了与本教材配套的PPT教学课件，包括各章全部课堂教学内容，以方便教师授课和学生自学。本书还新附了相应课程的教学基本要求和中英文名词索引，汇集了书中主要专业术语，以供参考查阅。

三、为真正体现以学生为中心的教材编写理念，全面提升学生的自主学习的能力，本次再版在每章之首增加简明的学习目标；每章正文内容以医药应用的案例引导贯穿教学全程；每章各节附有相关统计历史、统计学家简介和内容拓展等链接，以开阔学生视野和知识面；每章之后增加以简表形式高度概括的本章小结，再配以题型多样的目标检测及答案和统计软件应用的上机实训题等，从而有效帮助学生消化、巩固所学内容，提高其学习效率和成绩，全面提升其学习、实践和应用统计的能力。

本书仍由主编高祖新、尹勤主持全书的再版修订和统稿纂定工作，包括各章学习目标、本章小结等的编写。副主编徐伟主持了配套PPT教学课件的制作工作，副主编徐宁协助进行了全书的统稿审订工作，各章节的编委见各章结尾处。本书编著时注意博采众长，参考了国内外多种教材和参考文献，同时还得到科学出版社、编委所在单位及广大读者的大力支持、帮助和鼓励，在此一并表示衷心的感谢。

本书虽经反复认真修订，但由于水平有限，仍会有疏漏和不妥之处，恳请各位专家、读者继续批评指正，以便今后重新印刷或再版时修正完善。

编　者

2009年7月

第一版编写说明

为了适应我国高等职业教育的迅速发展，进一步深化高等职业教育的教学改革，培养高素质的医药高级技术应用人才，本教材作为医药高职高专系列教材之一，其编著既考虑到统计学科知识结构的科学性和系统性，又结合了医药生产和科研领域对统计应用的具体要求和特点，同时针对医药高职学生的接受能力和理解程度，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，内容系统而实用，写法上力求简明易懂，深入浅出，便于掌握。其主要特点是：

1. 作为基础课教材，本教材从高职高专各专业的培养目标和特点出发，在尽量保持统计学科的科学性和系统性的前提下，以“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”的目的性和“必需够用”的尺度。

2. 所选内容涵盖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，以统计数据的处理和分析为核心，把概率的基本理论融合到统计方法中去，注重统计方法思想的阐述，结合实际数据和实例说明统计方法的特点、应用条件和场合等，从而形成以统计原理、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。

3. 强化计算机应用的统计技能的培养。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，根据高职学生所用软件应满足普及、简明（不可太专业）的要求，本教材选用了Office 办公系统的 Excel 软件统计模块来进行统计软件应用的教学，不仅可减轻学生的计算负担，还可以提高其运用统计方法分析和解决实际问题的能力，真正达到“学以致用”的目的。

教材的主要内容有数据的描述和概括、概率论基础、抽样分布、参数估计、参数的假设检验、方差分析、正交设计、相关与回归分析等章节，并在每章的最后一节给出 Excel 软件对应统计功能的操作应用，同时辅之以相应的实例、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。

本教材供一学期使用，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，则可根据课时等客观条件来灵活取舍或安排学生自学。

本教材共分 8 章，由高祖新、尹勤主编而成。其中第 1～3 章、第 7 章和各章 Excel 软件应用部分由高祖新编写，第 4～6 章、第 8 章由尹勤编写，最后共同统稿纂定。

本教材的编著，得到了中国药科大学高职学院杨静化教授等专家的指导和帮助，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，书中不妥之处在所难免，恳请各位专家、读者批评指正。

编　者

2004 年 7 月

目　录

绪　论

医药统计是应用概率论与数理统计的原理和方法，对医药、生物等相关领域研究对象的数据资料信息进行搜集、整理、分析和解释，以显示其总体特征和统计规律性的应用科学。其中概率论（probability）是从数量侧面来研究随机现象统计规律性的数学学科，而数理统计（mathematical statistics）则是以概率论为基础，通过对随机现象观察数据的收集整理和分析推断来研究其统计规律的学科。

目前，我们所从事的医药研究和生产中，无论是疾病防治、药物研发、临床试验、公共卫生等各领域，还是新药研制、药物鉴定、药理分析、试验设计、药政管理、处方筛选、医药信息等医药领域的各个方面，都需要进行大量的数据资料的整理和分析，医药统计作为利用相关数据资料进行医药科学研究的重要前提和手段，其理论方法及应用已广泛渗透到医药研究与实践的各个领域，正起着越来越重要的作用。而有关医药统计的知识、方法和必要的统计软件应用技能训练，也已成为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1节　统计学及其发展简史

在日常生活中，统计既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；还可指分析统计数据的方法和技术，即统计学。统计学（statistics）是对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。

统计实践作为一种社会实践活动已有四五千年的历史，早在人类社会的初期——还没有文字的原始社会，就有了“结绳记事”等统计计数活动。但是，将统计实践上升到理论，使之成为一门系统科学——统计学，距今只有300多年的历史。最初的统计方法是随着社会政治和经济的需要而逐步得到发展的，直到18世纪概率论被引进之后，统计才逐渐形成一门成熟的科学。

最早的概率论萌芽之作是意大利数学怪杰卡尔达诺（G. Gardano，1501～1576年）于1563年撰写的《游戏机遇的学说》，书中讨论了两人赌博中断后分赌本问题，并提出了“大数定律”的基本概率理论的原始模型。到17世纪中叶，法国数学家帕斯卡（B. Pascal，1623～1662年）和费马（P. Fermat，1601～1665年）多次通信讨论了分赌本问题，并首次给出了正确答案。同时代的荷兰物理学家、数学家惠更斯（C. Huygens，1629～1695年）对此问题进行了深入研究，并于1657年出版了概率论的名著《论赌博中的计算》，书中提出了数学期望、概率的加法定理与乘法定理等基本概念。瑞士数学家雅科布·伯努利（Jakob Bernoulli，1655～1705年）创立了最早的大数定理——伯努利定理，建立了描述独立重复试验序列的“伯努利概型”，并撰写了最早的概率论专著——《猜度术》，使概率论成为一个独立的数学分支。

1662年英国统计学家格朗特（J. Graunt，1620～1674年）基于伦敦死亡人数资料的研究所进行的死亡率推算，是历史上最早出现的统计推断，他还发表专著《从自然和政治方面观察死亡统计表》，对人口统计、保险统计和经济统计进行了数学研究。1763年，英国统计学家贝叶斯（T. Bayes，1702～1761年）发表《论机会学说问题的求解》，给出“贝叶斯定理”，从结果去对原因进行后验概率的计算，可视为最早的数学化的统计推断。而最早将古典概率论引进统计学领域

的是法国天文学家、数学家拉普拉斯(P. S. Laplace,1749～1827 年),他提出了研究随机现象的分析方法,完善了古典概率论的结构,并阐明了统计学大数法则,进行了大样本推断的尝试。19 世纪初,德国著名数学家高斯(F. Gauss, 1777～1855 年)和勒让德建立"最小二乘法",且用之于分析天文观测的误差,高斯还成功地将正态分布理论用于描述观察误差的分布,并用于行星轨迹的预测。比利时统计学家凯特勒(A. Quetelet,1796～1874 年)发现了大量随机现象的统计规律性,开创性地应用了许多统计方法,并应用于天文、数学、气象、物理、生物和社会学等领域,完成了统计学和概率论的结合。此后,以概率论为基础的统计理论和方法被称为数理统计。

从 19 世纪中叶到 20 世纪中叶,数理统计和应用得到蓬勃发展并达到成熟。法国医生路易斯(P. C. A. Louis,1787～1872 年)研究了当时流行的用"放血"疗法治疗伤寒和肺炎效果,1835 年提出了医学观察中的抽样误差和混杂概念、临床疗效对比的前瞻性原则和疗效比较的"数量化"方法,被誉为"临床统计之父"。盖瓦勒特(J. Gavarret,1808～1890 年)1840 年在巴黎出版了世界上第一部医药统计教科书——《医学统计学》。德国的大地测量学者赫尔梅特(F. Helmert,1843～1917 年)在 1876 年研究正态总体的样本方差时,发现了 χ^2 分布(卡方分布)。英国生物学家、人类学家高尔顿(F. Galton,1822～1911 年)将正态分布理论用于社会学方面的研究,并在生物遗传学中提出了著名的回归、相关等概念,创立了回归分析法。数理统计学的奠基人之一、英国数学家、统计学家皮尔逊(K. Pearson,1857～1936 年)进一步发展了回归与相关的理论,提出了总体、标准差、正态曲线等重要术语和矩估计法、χ^2 拟合优度检验法,并创建了生物统计学,为 20 世纪数理统计和生物统计学的发展奠定了基础。英国统计学家戈塞特(W. S. Gosset,1876～1937 年)在 1908 年以"Student"为笔名在《生物统计学》杂志上发表论文,最早提出 t 统计量的精确分布——t 分布,开创了小样本统计理论的先河。而英国统计学派的代表人物费歇尔(R. Fisher,1890～1962 年)系统地发展了抽样分布理论,建立了以最大似然估计法为中心的点估计理论,首创了试验设计法并提出方差分析法,奠定了统计学沿用至今的数学框架,被誉为现代数理统计学的奠基人之一。1933 年原苏联的著名数学家柯尔莫哥洛夫(A. N. Kolmogrov,1903～1987 年)出版了经典名著《概率论基础》,首次以测度论为基础建立了概率的公理化定义,从而使概率论建立在完全严格的数学基础之上,奠定了现代概率论的理论基础。美国统计学家奈曼(J. Neyman, 1894～1981 年)和小皮尔逊(E. Pearson,1895～1980 年,K. Pearson 之子)合作,20 世纪 30 年代提出了似然比检验,并建立了置信区间理论,在数学上完善了假设检验和区间估计的理论体系。而美籍罗马尼亚统计学家沃尔德(A. Wald,1902～1950 年)所建立的序贯分析和统计决策理论,美国统计学家威尔克斯(S. Wilks,1906～1964 年)所创立的多元方差分析、多项式分布、多变量容许区间等一系列多元分析方法,开创了数理统计学的新局面。1946 年瑞典数学家克拉默(H. Cramer,1893～1985 年)发表《统计学的数学方法》,运用测度论方法总结数理统计的成果,使现代数理统计趋于成熟。

随着自然科学和社会经济的进步和发展,数理统计在理论上不断成熟与完善,应用上日益广泛和深入。数理统计也成为研究自然现象和社会经济现象数量方面的极为有力的工具,并逐步渗透到各个学科领域,形成了许多边缘学科,如信息论、决策论、排队论、可靠性理论、自动控制、统计质量管理、生物统计、医药统计、社会统计、水文统计、统计物理学、计量经济学、计量心理学等,成为现代科学发展的一个重要标志。

第 2 节　常用统计软件简介

随着电子计算机的应用和普及,特别是计算机统计软件的深入发展,人们的数据处理能力大大增强,以往受计算能力限制的数理统计有关理论和方法,其处理实际问题的能力也得到了

空前提高。统计软件是利用计算机软件技术呈现统计数据，进行数据分析，模拟和实现统计过程的一类专业应用软件，是统计方法应用的重要载体，在医药统计数据处理和统计分析中具有日益重要的地位。

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、统计图表显示和统计分析等工作。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

SAS 系统，全称 Statistical Analysis System（统计分析系统），是模块化、集成化的大型应用软件系统，具有完备的数据管理、数据分析、数据存取、数据显示等功能，在数据处理方法和统计分析领域，被誉为国际上的标准软件和最具权威的优秀统计软件包。

SAS 系统最初是由美国北卡罗来纳州州立大学的 A. J. Barr 和 J. H. Goodnight 教授于 20 世纪 60 年代末期开始研发的，1975 年在美国创建 SAS 研究所（SAS Institute Inc.），之后推出的 SAS 系统 SAS/PC、SAS for Windows 等版本始终以领先的技术和可靠的支持著称于世，并不断发展与完善。SAS 系统中提供的主要分析功能包括统计分析、经济计量分析、时间序列分析、决策分析、财务分析、全面质量管理、运筹规划、地理信息系统分析和医药临床研究等，已广泛应用于自然科学、社会科学、经济管理、医药研究等各领域，为全球 100 多国家和地区的众多用户所采用，是当今国际上最著名的数据分析软件系统。

SAS 系统是一个组合的软件系统，它由多个功能模块配合而成，其基本部分模块是 BASE SAS（SAS 基本模块），还可以根据需要增加 SAS/STAT（统计分析模块）、SAS/GRAPH（绘图模块）、SAS/QC（质量控制模块）、SAS/ETS（经济计量学和时间序列分析模块）、SAS/OR（运筹学模块）、SAS/IML（交互式矩阵程序设计语言模块）、SAS/FSP（快速数据处理模块）、SAS/AF（交互式全屏幕软件应用系统模块）、SAS/GIS（地理信息系统模块）、SAS/MAP（地图模块）、SAS/MDDB（多维数据处理模块）等十多个模块进行组合来增加不同的功能。

SAS 提供的统计分析软件，覆盖了所有实用的数理统计分析方法，包括回归分析、方差分析、相关分析、主成分分析、因子分析、多元分析、聚类分析、判别分析、分类数据分析、图表分析等多个统计分析过程，每个过程均含有极丰富的任选项。SAS 提供的绘图系统，不仅能绘各种统计图，还能绘出地图。

然而，由于 SAS 系统是从大型机上的系统发展而来的，其全面的操作仍以编程为主，系统地学习掌握需要花费较多的精力。

二、SPSS

SPSS，原名全称 Statistical Package for the Social Science（社会科学统计软件包），2000 年 SPSS 公司将其英文全称改为“Statistical Product and Service Solutions”（统计产品与服务解决方案），是集数据整理、分析功能于一身的组合式大型通用统计分析软件包，以其强大的统计分析功能、方便易用的用户操作方式、灵活的表格分析报告和精美的图形展现形式，与 SAS 同为当前世界上最为流行的应用最广泛的专业统计分析软件。

SPSS 最早是由美国斯坦福大学的三位研究生于 20 世纪 60 年代末研制开发，同时还成立了 SPSS 公司。1984 年 SPSS 公司推出了世界第一套统计分析软件微机版本 SPSS/PC +，开创了 SPSS 微机系列产品的先河。目前 SPSS 已推出 9 个语种版本，不仅应用于社会科学领域，而且广

泛应用于自然科学、商务经济、医药卫生、政府部门、教学科研等各个领域。世界上许多有影响的报纸杂志纷纷就 SPSS 的自动统计绘图、数据深入分析、使用灵活方便、功能设计齐全等方面给予了高度的评价。目前的 SPSS for Windows 版本，使用 Windows 的窗口方式展示各种管理和分析数据方法的功能，使用对话框展示出各种功能选择项，只要掌握一定的 Windows 操作技能，粗通统计分析原理，就可以使用该软件进行各种数据分析，因此深受广大应用统计分析人员的欢迎，目前在国内也已广泛流行起来。

SPSS for Windows 是模块结构的组合式软件包，它集数据整理、分析功能于一身，用户可以根据实际需要和计算机的功能选择模块。其模块主要有：SPSS Base、SPSS Advance、SPSS Categories、SPSS Complex Sample、SPSS Exact Test、SPSS Maps、SPSS Regression 、SPSS Table 和 SPSS Trends 等十多个。SPSS 的基本功能包括数据管理、统计分析、图表分析、输出管理等。其统计分析(analyze)过程包括描述性统计、均值比较分析、一般线性模型、相关分析、方差分析、回归分析、非参数检验、主成分分析与因子分析、对数线性模型、聚类分析与判别分析、数据简化分析、生存分析、时间序列分析、多重响应变量分析等几大类，每类中又分好几个统计过程，如回归分析中又分线性回归分析、曲线估计、Logistic 回归、Probit 回归、加权估计、两阶段最小二乘法、非线性回归等多个统计过程，而且每个过程中又允许用户选择不同的方法及参数。SPSS 也有专门的绘图系统(graph)，可以根据数据绘制各种统计图形和地图。同时 SPSS 可以直接读取 Excel 及 DBF 数据文件，并已推广到多种操作系统的计算机上，全面适应互联网。

方便易用是 SPSS for Windows 的主要优点，同时也是 SPSS 不够全面的原因所在。

三、Excel

Excel 作为 Microsoft Office 办公软件包的最重要的组件之一，是一个功能强大且使用简便的电子表格软件，它方便实用、界面友好、系统自带、极为普及。它的基本职能是对数据进行记录、计算与分析，不仅具有强大的制表和绘图功能，而且内置了数学、统计、财务等 10 类 300 多种函数，同时还提供数据分析、规划求解、方案管理器等多种分析方法和工具，可进行各种数据处理、基本统计分析、数学计算和辅助决策操作等。

与标准的专业统计分析软件 SAS、SPSS 等相比，Excel 虽然其统计分析的范围和内容没有 SAS、SPSS 等全面深入，但其优势也很明显，主要在于其强大的数据自动填充功能，使数据输入变得相当简单；方便的数据汇总与数据透视分析功能，可快速得到整个工作表及有关探索性统计分析结果；完美的图表内置格式，使得统计制作方便，图形美观；而 Excel 作为一般计算机内都配置的 Microsoft Office 办公软件的重要组件，且为中文界面，通常无版权问题，其使用的方便普及是其他统计专业软件所无法相比的。

Excel 有很强的进行数据基本统计处理的功能。它提供了一组数据分析工具，称为“分析工具库”，在建立复杂的统计分析时，使用现成的数据分析工具很便捷，只需为每一个分析工具提供必要的数据和参数，该工具就使用适宜的统计或数学函数及内置程序，在输出表格中显示相应的结果。利用“分析工具库”所生成的“数据分析”工具，Excel 可完成常用的基本统计分析，包括各种描述性统计、数据库统计函数与数据透视表、统计指数、概率分布、抽样与抽样分布、参数估计、假设检验、非参数检验、方差分析、相关分析、回归分析和预测、时间序列分析等内容。

由于 Excel 软件普及程度高，操作运算也较为简便，本书将结合各章的内容，介绍 Excel 软件相应的统计分析与运算处理操作应用，以提高和拓展数据处理和统计分析的应用能力。

“统计”词语的产生

统计已经有几千年的历史，不过在早期还没有出现“统计”这样的用语。

统计词源最早出现于中世纪拉丁语的“Status”，指各种现象的状态和状况。由这一语根组成意大利语“Stato”，表示“国家”的概念，也含有国家结构和国情知识的意思。而最早作为学名使用的“统计”，是在18世纪德国政治学教授亨瓦尔（G. Achenwall）在1749年所著的《近代欧洲各国国家学纲要》绪言中，把国家学名定为“Statistika”（统计）这个词。原意是指“国家显著事项的比较和记述”或“国势学”，认为统计是关于国家应注意事项的学问。此后，各国相继沿用“统计”这个词，并把这个词译成各国的文字，法国译为“Statistique”，意大利译为“Statlstica”，英国译为“Statistics”，日本最初译为“政表”、“政算”、“国势”、“形势”等，直到1880年在太政官中设立了统计院，才确定以“统计”二字正名。

1903年（清光绪二十九年）由钮永建、林卓南等翻译了4本横山雅南所著的《统计讲义录》，把“统计”这个词从日本引入我国。1907年（清光绪三十三年）彭祖植编写的《统计学》在日本出版，同时在国内发行，这是我国最早的一本“统计学”书籍。“统计”一词就成了记述国家和社会状况的数量关系的总称。

链接

（高祖新）

第1章　数据的描述和统计概括

学习目标

1. 掌握数据的类型及特性
2. 掌握定性和定量数据的整理步骤、显示方法
3. 了解统计图形和统计表的表示及意义
4. 掌握描述数据分布的集中趋势、离散程度的常用统计量
5. 能理解并熟练掌握样本均值、样本方差的计算
6. 了解用 Excel 软件进行统计作图、频数分布表与直方图生成、统计量的计算

统计学是对研究对象的数据资料进行搜集、整理、分析和研究，以显示其总体的特征和规律性的学科。统计学的研究对象是客观事物的数量特征和数据资料。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资料，可见，统计学与统计数据是密不可分的。

【案例 1-1】

根据《中国人口统计年鉴 2001》提供的 2000 年我国人口普查数据资料，在我国 6 周岁以上人口中按不同文化程度分为：文盲半文盲、小学、初中、高中及中专、大专及以上等 5 组，其中 1.1093 亿人是文盲半文盲；4.5191 亿人是小学文化程度；4.2989 亿人是初中文化程度；1.4109 亿人是高中及中专文化程度；0.4571 亿人是大专及以上文化程度。

问题：

如何对上述文化程度资料进行统计整理，并用统计图表显示？

【案例 1-2】

现有某高校某专业 110 名学生统计课程的成绩（分）数据如下

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

问题：

（1）该成绩数据与案例 1-1 的文化程度资料有何区别？

（2）如何对该成绩数据进行统计整理，并用统计图表显示？

本章我们就讨论如上述案例所示的有关数据资料的统计整理、图表显示和统计概括等问题。

第 1 节　数据的类型和整理

一、数据的类型

数据(data)或资料是对客观现象计量的结果。例如,对药品质量的计量可得到药品是正品或次品的数据;对药物在试验对象血液中含量的计量可得到血液浓度数据等。统计数据是利用统计方法进行分析的基础,不同的统计数据应采用不同的统计分析方法。

(一) 数据的类型

由于对事物计量的精确程度不同,我们可以得到不同类型的数据,需用不同的统计分析方法进行分析处理。实际统计应用中,对应于不同的计量尺度,数据可分为定类数据(或名义数据、计数数据)、定序数据(或有序数据、等级数据)和数值型数据(或计量数据)三种类型。

1. 定类数据(categorical data 或名义数据 nominal data、计数数据 count data)　是对事物按照其属性进行分类或分组的计量结果,其数据表现为文字型的无序类别,可以进行每一类别出现频数的计算,但不能进行排序和加减乘除的数学运算。例如,人口的性别分为男、女两类;人体血型分为 O 型、A 型、B 型和 AB 型四类等,这些均属于定类数据。

2. 定序数据(ordinal data 或有序数据、等级数据 rank data)　是对事物之间等级或顺序差别的计量结果,其数据表现为有序类别,可以进行类别的频数计算和排序,但不能进行加减乘除的数学运算。例如,某种药物的疗效可分为“有效”、“无效”两类;考试的等级成绩可分为优、良、中、及格、不及格这五类等,均属于定序数据。案例 1-1 中的我国人口的文化程度数据也属于定序数据。

3. 数值数据(numerical data 或计量数据 measurement data)　是用自然或度量衡单位对事物进行计量的结果,其数据表现为具体的数值,既可进行频数计算和排序,又可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;人的体重、血压、红细胞数等,均为定量数据。

定类数据和定序数据说明的是事物的品质特征,其结果通常表现为类别,故可统称为定性数据(qualitative data)或品质数据。数值数据说明的是事物的数量特征,是用数值来表示的,其结果通常表现为量化的具体数字,故又称为定量数据(quantitative data)。

(二) 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(variable value)。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(categorical variable 或名义变量 nominal variable)、定序变量(ordinal variable 或等级变量 rank variable)和数值变量(numerical variable 或 metric variable)。

数值变量中,如果变量可以取有限个值或可列无穷多个数值,即可以一一列举,称为离散变量(discrete variable),如制药公司个数、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、血药浓度等。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。

区分数据的类型非常重要，如表 1-1 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 1-1 不同数据类型之比较

数据类型	定性数据(品质数据)		定量数据
	定类数据 (计数数据)	定序数据 (等级数据)	数值数据 (计量数据)
表现形式	类别 (无序)	类别 (有序)	数值 (+ − × ÷)
对应变量	定类变量	定序变量	数值变量 (离散变量、连续变量)
主要统计方法	计算各组频数，进行列联表分析、χ^2 检验等非参数方法		计算各种统计量，进行参数估计和检验、回归分析、方差分析等参数方法

(三) 两类数据的转换

根据统计分析的需要，定量数据与定性数据之间经常要做数据类型的转换。

1. 定量数据的定性化转换 例如，作为定量数据的成年男子的血清胆固醇值，按是否小于 6(mmol/L)划分成血脂正常和异常两类，就转化为定性数据。若将血红蛋白按含量(g/L)的多少分为五级：<60(重度贫血)、60～90(中度贫血)、90～120(轻度贫血)、120～160(血红蛋白正常)、>160(血红蛋白增高)，这时定量数据就化成了定性数据。

2. 定性数据的数量化转换 为了便于统计处理，我们有时需要对定性数据赋值进行数量化转换。例如，对定性变量性别中的定性数据“男”“女”可以分别取值为“1”和“0”，此时取值 1 和 0 之间没有量的差别，只是一种“数据代码”。又如对文化程度，如果是按文盲半文盲、小学、初中、高中、大学及以上这 5 组进行分类，则文化程度变量属于定序变量，对这 5 类数据赋值时我们可分别取值为 1、2、3、4、5，此时取值 1、2、3、4、5 之间不仅是一种“数据代码”，也有量的区别。

(四) 统计数据的搜集和来源

统计数据资料的搜集是指根据统计研究的目的，采用科学研究或调查方法，向研究或调查对象搜集数据的过程，它是统计分析的基础。统计数据资料搜集的基本要求是：准确性、及时性和系统性。通过数据搜集，我们可得到两类不同来源的数据资料：

1. 原始资料(primary data 或一手资料) 通过专门进行的科学试验或调查来采集得到的直接来源数据资料。其中科学试验是取得自然科学数据的重要手段，而专门调查是取得社会经济数据的重要手段。

2. 次级资料(secondary data 或二手资料) 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的间接来源数据资料，包括图书资料和报纸杂志、广播电视等媒体和因特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。

二、数据的统计整理和图示

数据的统计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汇总和处理，

使数据资料系统化，以反映研究总体的特征、规律和趋势。数据整理和图示通常包括数据的审核筛选、分类或分组、汇总、给出统计图表或报告等步骤。

在对数据进行统计整理时，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，对定性数据（定类数据和定序数据）主要作分类整理，对定量数据（数值数据）主要作分组整理。

（一）定性数据的整理和图示

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或类别排序，进行数据整理时，只需按不同数据（类别）进行分组，算出各组的频数或频率、百分比（对于定序数据，还可以算出各组的累积频数或累积频率、累积百分比），列出频数分布表，再用条形图或圆形图等统计图形显示其整理结果。

频数（frequence 或 frequency）是指落在各类别中的数据个数；频率（frequency 或 relative frequency）则是指各类别的数据个数占数据总个数的比例值；我们将各个类别及其相应的频数（或频率、百分比）用表格形式全部列出来就是频数分布表（frequency table）。

【案例 1-1】

解：根据 2000 年我国人口普查数据中我国 6 周岁以上不同文化程度的数据资料，就可得到下列频数分布表。

表 1-2　2000 年我国 6 周岁以上各种文化程度的人口数

文化程度	人数（亿）	百分比（%）
文盲半文盲	1.1093	9.4
小学	4.5191	38.3
初中	4.2989	36.4
高中及中专	1.4109	12.0
大专及以上	0.4571	3.9
合计	11.7953	100.0

资料来源：国家统计局编《中国人口统计年鉴 2001》，中国统计出版社。

利用表 1-2 的数据，我们就可作出 2000 年我国各种文化程度人口数的（垂直）条形图，见图 1-1，它直观地反映了我国各种文化程度的人口分布状态。

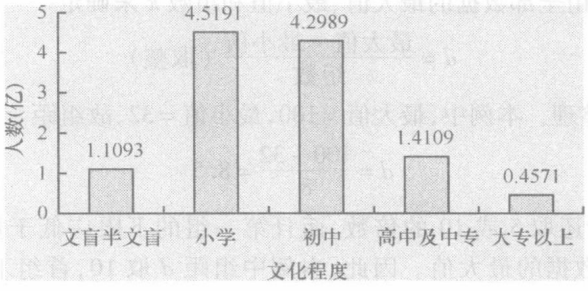


图 1-1　2000 年我国 6 周岁以上人口的各种文化程度的垂直条形图

对定性数据或离散变量数据，条形图和圆形图（本章第 2 节）是反映数据分布特征和构成比的常用统计图形，在统计图表显示中起着很好的作用，我们将在本章第 2 节简要介绍这两种统计图形。

（二）定量数据的整理和图示

对于定量数据（数值数据）主要作分组整理。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、离散趋势及形状等统计特征。其整理和图示主要包括按数量标志进行分组，编制频数分布表，并采用直方图及频数折线图等统计图形来表示其整理结果，以更直观清晰地表示其频数分布状态。

定量数据统计分组方法有单变量值分组和组距分组两种。单变量值分组是按每个变量值作为一组，主要用于离散变量且变量值较少情形。对于连续变量或变量值较多情形，通常采用组距分组，即将全部变量值依次划分为若干个区间，每个区间作为一组。在组距分组中，一个组的最小值称为该组的下限（lower limit）、最大值称为该组的上限（upper limit）。

下面我们结合前面提出的案例 1-2 介绍组距分组法编制频数分布表的方法。

【案例 1-2】

解：（1）显然，该成绩数据是定量数据，而案例 1-1 的文化程度数据是定性数据中的定序数据，是属于不同类型的数据。

（2）下面我们结合该成绩数据的整理和图示，给出定量数据组距分组法编制频数分布表步骤。

1. 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 或 6 组；当数据个数为 100 左右时，可分为 6 ~ 10 组；当数据个数超过 500 时，可分为 10 ~ 15 组。在实际分组时，也可按斯塔基（Sturges）提出的经验公式来定组数 k

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

式中，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。例如，在本例中，$N = 110$，则

$$k = 1 + \frac{\ln 110}{\ln 2} = 7.78$$

即大致可分为 8 组。

2. 确定组距　在分组中，组距（class width）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 来确定

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

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，组距有时还取 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。

3. 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（如用 Excel 软件，参见本章第 4 节），计算各组频数，即可列出频数分布表，见表 1-3。