中国科学院教材建设专家委员会规划教材
全国高等医药院校规划教材

供临床、基础、口腔、麻醉、影像、药学、检验、护理、法医等专业使用

案例版™

医学统计学

第3版

主　编　罗家洪　郭秀花

科学出版社

中国科学院教材建设专家委员会规划教材
全国高等医药院校规划教材

案例版 TM

供临床、基础、口腔、麻醉、影像、药学、检验、护理、法医等专业使用

医学统计学

第 3 版

主　　编　罗家洪　郭秀花
副 主 编　姚应水　贾　红　刘启贵　赵若望　董莉萍　程晓萍　李晓梅　李秀央　谢红卫　罗艳侠
学术秘书　毛　勇　彭林珍
编　　委　（按姓氏笔画排序）

王良君（锦州医科大学）
毛　勇（昆明医科大学）
刘军祥（西南医科大学）
刘　艳（南华大学）
李晓梅（昆明医科大学）
肖媛媛（昆明医科大学）
吴梦吟（浙江大学）
宋桂荣（大连医科大学）
陈　莹（昆明医科大学）
罗　健（昆明医科大学）
和丽梅（昆明医科大学）
赵若望（内蒙古科技大学包头医学院）
胡志宏（北华大学）
俞婉琦（浙江大学）
贺连平（皖南医学院）
郭秀花（首都医科大学）
康耀文（皖南医学院）
董莉萍（北华大学）
程晓萍（锦州医科大学）
谢红卫（南华大学）
王耶盈（昆明医科大学）
叶运莉（西南医科大学）
刘启贵（大连医科大学）
李秀央（浙江大学）
李　霞（首都医科大学）
吴立娟（首都医科大学）
何利平（昆明医科大学）
张俊辉（西南医科大学）
罗艳侠（首都医科大学）
罗家洪（昆明医科大学）
孟　琼（昆明医科大学）
郝金奇（内蒙古科技大学包头医学院）
侯瑞丽（内蒙古科技大学包头医学院）
姚应水（皖南医学院）
贾　红（西南医科大学）
常　巍（昆明医科大学）
彭林珍（云南交通职业技术学院）
喻　箴（昆明医科大学）
童玲玲（南华大学）
詹志鹏（锦州医科大学）

科学出版社
北　京

郑 重 声 明

为顺应教育部教学改革潮流和改进现有的教学模式，适应目前高等医学院校的教育现状，提高医学教育质量，培养具有创新精神和创新能力的医学人才，科学出版社在充分调研的基础上，引进国外先进的教学模式，独创案例与教学内容相结合的编写形式，组织编写了国内首套引领医学教育发展趋势的案例版教材。案例教学在医学教育中，是培养高素质、创新型和实用型医学人才的有效途径。

图书在版编目（CIP）数据

医学统计学 / 罗家洪,郭秀花主编. —3 版. —北京:科学出版社，2018.3
中国科学院教材建设专家委员会规划教材·全国高等医药院校规划教材
ISBN 978-7-03-056283-8

Ⅰ. ①医… Ⅱ. ①罗… ②郭… Ⅲ. ①医学统计-医学院校-教材
Ⅳ. ①R195.1

中国版本图书馆 CIP 数据核字（2018）第 006872 号

责任编辑：朱 华 / 责任校对：郭瑞芝
责任印制：赵 博 / 封面设计：王 融

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
新科印刷有限公司 印刷
科学出版社发行 各地新华书店经销
*
2006 年 8 月第 一 版 开本：850×1168 1/16
2018 年 3 月第 三 版 印张：17 1/2
2018 年 3 月第二十三次印刷 字数：580 000
定价：68.00 元
（如有印装质量问题，我社负责调换）

前　言

本教材是根据《国家中长期教育改革和发展规划纲要（2010—2020年）》、《国家中长期人才发展规划纲要（2010—2020年）》《国家中长期科学和技术发展规划纲要（2006—2020年）》等的精神，在本科案例版《医学统计学》第2版的基础上编写的第3版教材，本教材本着与时俱进、改革与创新医学生培养模式、教学方法的宗旨，在借鉴国外先进教学模式——案例式教学模式的基础上，编写的适合中国国情的全新案例式教材。

医学统计学是医学生的一门重要基础学科，它是应用统计学的基本原理和方法，研究医学及其有关领域数据信息的搜集、整理、分析、表达和解释的一门科学。由于医学领域及其有关领域的研究主要对象是人，人的健康及其影响因素较复杂，具有生物变异性和多因素特点，与社会因素、心理因素、环境因素等有关，需要借助医学统计学的方法进行统计分析，解决医学日常工作和医学科研工作的实际问题。因此，医学统计学是医学生的公共基础课，又是专业基础课；医学统计学是医疗卫生人员正确认识医学领域及其有关领域的客观规律、总结工作经验、进行医学科研和疾病防治工作的重要工具。人类已进入信息时代，要在大量的信息中获得有价值的结果，需要对信息进行科学的统计分析，这就要求医学生具有扎实的医学统计学基础。为培养合格的21世纪医学人才，绝大多数医科院校都将医学统计学列为本专科生、硕士研究生、博士研究生以及医学继续教育和成人在职培训的必修课程。但是，由于医学统计学的抽象性、综合性和灵活性等特点，加之传统的教材内容和教学模式，使得医学生普遍感到该课程难学、难懂，以至于在医疗实践中不知如何应用。

长期以来，我国高等教育的教学活动中“教”“学”分离现象突出，枯燥的“填鸭式”教学，单向传输的师生关系，导致医学生学习主动性不够，创新思维不强，自学能力缺乏，影响了人才培养的质量。相对于其他学科而言，医学教学模式更为传统和保守，课程体系、教学方法几十年不变，这与“灌输式”的教材结构有着很大关系。传统教材模式为基本概念→基本理论→正常案例→习题，没有错误案例，学生不知道会犯何种错误？应该怎样避免？怎样正确分析？为了顺应教育部教学改革的潮流，改进现有教学模式和课程体系，提升教学质量和就业率，我们在不改变教学核心内容的前提下，引进国外先进教学模式，借鉴职业教育成功经验，以全新面貌编写了案例式医学统计学，其主要特点有以下几个方面：

1. 先进性　在突出基础理论、基本知识和基本技能的基础上，融典型科研正反案例于教材中，以案例引导教学，采用错误案例（或正常案例）→问题→分析→引导出基本概念→基本理论→实际科研案例→正确分析方法→知识点→思考练习模式，丰富教学内容，提高学习效率。强调以“学”为中心，以学生的主动学习为主，打破传统教学中强调的以“教”为主，将教学改革落到实处。

2. 科学性　注重创新能力和实践能力的培养，力求为学生知识、素质和能力的协调发展创造条件。将教学改革和教学经验、临床科研成果融入教材，基础课程中加入临床案例，加强了基础学科与临床学科的联系和结合，明确了学习基础课的目的，让学生感到学有所用，既能充分调动学习主动性和积极性，提高学习效率，又能大幅度提升教学质量。

3. 启发性　用各种正确和错误的典型案例启发学生思考，引导学生提出问题，鼓励学生自己寻找问题的答案，培养学生批判性和分析性的思维能力，从根本上改变死记硬背、理论与实践相脱离的学习过程。

4. 实用性　各章节知识点明确，学生易学，教师好教，使学生在较短的时间内掌握所学知识。教材内容符合教育部制定的基本教学要求，以5年制医学本科生为主要对象，以临床医学专业为主，兼顾预防、基础、口腔、影像、麻醉、护理、药学、检验、视光、社保等专业需求，同时适用于医学生全国统考、毕业后执业医师考试和硕士研究生入学考试，也可作为在职医疗卫生人员继续教育培训教材，还可以作为在职医疗卫生人员科研参考书。

5. 激励或行动指南　每章有许多名人名言，在学习《医学统计学》时也熟悉许多名人名言，这些名人名言将给予读者激励、警戒或作为行动的指南。

本教材第1版出版后受到读者的一致好评，获得2010年云南省优秀教材和昆明医学院优秀教材。第2版在第1版的基础上，增加了二项分布与Poisson分布、多因素分析两章内容，每章新增许多名人名言，给读者激励、警戒或作为行动的指南。第3版在第2版基础上，更换了最新科研案例，修改了第2版不足之处，增添了许多思考练习题。

本教材是常年从事医学统计学和卫生统计学教学工作的各位主编、副主编及编委的经验总结，也是医学科研统计方法的综合反映。在教材编写和出版过程中，得到了科学出版社医学分社社长李国红及朱华编辑等和各参编医科院校的大力支持；同时，昆明医科大学校长李松教授，副校长李燕主任医师，公共卫生学院院长殷建忠教授、罗勇前书记、许传志副院长等也给予了大力支持并提出了宝贵意见，我谨代表全体编委一并鸣谢。

本教材是全新的案例式教材，限于我们的水平和编写经验，可能有不少的缺点和错误，热忱欢迎广大师生和同行批评指正，并希望各医科院校在使用过程不断总结经验，提出宝贵意见，以便进一步修改提高。

罗家洪

2017 年 11 月于春城昆明

目　录

第 1 章　绪　论

第一节　概　述

【例 1-1】

某医师研究中西药治疗胃溃疡患者疗效，在进行简单实验设计后，随机抽取胃溃疡患者 120 人作为研究对象，用随机方法将研究对象随机分成两组，分别采用中药和西药进行治疗，两组除用药不同外，其他条件尽可能相同；观察时采用盲法观察。中药组治疗 60 人，有效 54 人，有效率为 90.0%；西药组治疗 60 人，有效 42 人，治愈率为 70.0%，该医师认为中药治疗胃溃疡患者的疗效高于西药。该医师将资料整理撰写成论文，投稿到某杂志编辑部，没有几天，该医师接到该杂志编辑部的回信：请重新做统计学处理。该医师非常不理解，已经计算了各组的有效率，还要做什么统计处理？

【问题 1-1】

（1）该医师得到的资料属于何种类型资料？

（2）该资料属于何种设计方案？

（3）为什么杂志编辑部编辑要求重新做统计学处理？

（4）该资料需要用何种统计方法处理？

读一本好书，就是和许多高尚的人谈话。

——歌德

【分析】

（1）该资料的治疗结果按有效和无效分类，分别清点中西药组的有效和无效人数，属于典型的二分类计数资料。

（2）该医师采用随机抽样和随机分组方法，属于完全随机设计（成组设计）方案。

（3）该医师采用抽样研究，不可避免存在抽样误差，不能直接凭统计描述指标即有效率大小下结论，故该杂志编辑部编辑要求重新做统计学处理，需要进一步进行统计推断即假设检验后根据 P 值大小再下结论。

（4）根据资料的类型及其设计方案，应采用四格表 χ^2 检验。

该医师请统计学专家进行四格表 χ^2 检验，得 $\chi^2 = 7.500$，$P = 0.006$，小于检验水准 α（α=0.05），**中西药有效率差异有统计学意义，可认为中西药治疗胃溃疡患者的有效率有差别。**通过资料的统计分析，该医师深刻认识到：**没有较好的统计学知识，就不可能进行较好的科学研究，更不可能写出一篇高质量的科研论文。**

统计学（statistics）是应用概率论和数理统计的基本原理和方法，研究数据的搜集、整理、分析、表达和解释的一门科学。通过对被研究对象的数量信息的综合分析，去粗取精，去伪存真，透过表面现象揭示事物内部的客观规律。将统计学的理论和方法应用于自然科学和社会科学的不同领域，形成了若干统计学的分支，医学统计学就是其中之一。国际统计学界把医学统计学称为生物统计学（biostatistics）。

有志者，事竟成。

——［南朝］范晔

医学统计学（medical statistics）是应用统计学的基本原理和方法，研究医学及其有关领域数据信息的搜集、整理、分析、表达和解释的一门科学。由于医学领域及其有关领域的研究主要对象是人，人的健康及其影响因素较复杂，具有生物变异性和多因素特点，与社会因素、心理因素、环境因素等有关，需要借助医学统计学的方法进行统计分析，解决医学日常工作和医学科研工作的实际问题。因此，医学统计学是医学生的公共基础课，又是专业基础课；医学统计学是医疗卫生人员正确认识医学领域及其有关领域的客观规律、总结工作经验、进行医学科研和疾病防治工作的重要工具。

医学统计学的主要内容：①基本理论与方法：包括研究设计（调查设计、实验设计）、统计描述（计量资料的统计描述、分类资料的统计描述、统计图表）、统计推断（t 检验、方差分析、χ^2 检验、秩和检验等）、直线相关与回归、多元统计分析等。②健康统计：包括医学人口统计、疾病统计、健康体检统计等。③医疗服务统计：病案统计、医院统计、医疗服务的需求与利用、医疗保健制度与管理的统计分析等。本教材内容主要是医学统计学的基本理论和方法，有些教材上没有的方法可以参考其他统计教材。

统计学的起源可以追溯到18世纪甚至更早，但统计学的主要发展却是在19世纪末和20世纪初才真正开始，到40年代才逐渐成熟起来的。第二次世界大战以后，随着电子计算机的发展，统计的计算工作变得简化而快捷，计算工作全部用计算程序完成。将对资料进行各种统计处理分析的一系列程序组合就成为统计软件包（statistical package），有了统计软件包以后，统计学的计算非常简便，应用更为广泛。

统计软件包较多，国际上最著名的三大统计软件包为SAS、SPSS和BMDP，SAS（statistical analysis system，统计分析系统）是由美国的SAS institute在20世纪60年代开发的统计软件包。SPSS（statistical package for social sciences，社会科学统计软件包）是美国SPSS公司开发的大型统计软件包。自BMDP被SPSS收购后，第三大国际统计软件包一直没有确定。在国际学术界有条不成文的规定：凡是用SAS和SPSS统计分析的结果，在国际学术交流中可以不必说明算法。此外，国际统计软件包还有STATA、SYSTAT、STATISTICA、S-PLUS、P-STAT、SPIDA、MINITAB、GLIM、EPI INFO、R软件等；国内有四川大学华西公共卫生学院卫生统计学教研室开发的PEMS（The Package for Encyclopaedia of Medical Statistics，《中国医学百科全书·医学统计学》统计软件包）中文统计软件包等。目前，统计学及其软件包已广泛普及和应用于医学、社会学、市场学、经济学和自然科学等各个领域的信息处理、定量研究和科研分析中。

【知识点1-1】 统计学的概念

1. 统计学是应用概率论和数理统计的基本原理和方法，研究数据的搜集、整理、分析、表达和解释的一门科学。

2. 医学统计学是应用统计学的基本原理和方法，研究医学及其有关领域数据信息的搜集、整理、分析、表达和解释的一门科学。

3. 统计软件包是对资料进行各种统计处理分析的一系列程序的组合。

每一个成功者的秘诀，是由于坚定不移的志向和热烈不懈的工作。
——[英]马尔顿

医学生学习统计学，不仅多了一个认识世界、了解事物的工具，而且对将来的工作也有一定的帮助。医学工作和研究的对象是富于变化的人体，目前人类对自己的身体还有许多未知的部分，医学的发展离不开统计学的应用。对医生和医学科研工作者，阅读文献是了解学科发展的必不可少的工作，统计学知识有助于在阅读中对文献的可靠性进行正确的判断，也有助于更好地理解别人的工作。如果要进行科学实验或将自己的工作总结发表，更是离不开统计学知识。而统计学的基本思想，有助于形成良好的逻辑思维，增强科研工作能力。

学习统计学，不是像学数学那样单纯理论和做习题，也不是像学医学那样记忆许多细节，事事眼见为实。学习统计学必须理论课与实习课紧密结合。理论课着重掌握基本概念、基本思想与基本方法，掌握它们的意义、用途和应用条件，而不必深究其数学推导和计算；实习课通过讨论和案例分析，加深对“三基”的理解与认识，并通过一些应用分析题的解题练习，掌握医学统计学的分析过程和方法。

学习统计学的最好方法是熟练掌握一个统计软件包，如SPSS或SAS。根据实际资料正确选用分析方法，自己操作统计软件包完成有关计算，并对计算结果做出合理的解释。在实际操作过程中，可以加深统计方法的理论和应用条件的理解。学习统计是一个长期的过程，经常撰写科研论文有助于对所学知识的掌握和巩固，也有助于提高统计方法的应用能力。

第二节 统计工作的步骤

例1-1例子提示统计工作步骤分为研究设计→搜集资料→整理资料→分析资料→撰写科研论文，一般把前四步作为统计工作的基本步骤，简化为“**设计、搜集、整理、分析**”八字。

千里之行，始于足下。
——老子

1. 研究设计（research design） 科研结果的好坏取决于研究设计的好坏，一定的设计决定了一定的数据分析方法，不同设计方案下获得的资料要用不同的方法来分析，因此，设计决定分析，选择统计方法时应首先弄清楚资料类型属于何种资料，采用什么设计方案。设计是整个研究过程中最关键的一环，因此将在调查设计和实验设计中进行专门的介绍。

2. 搜集资料（data collection） 对统计资料的收集要做到完整、准确、及时、可靠。医学科学研究的

资料主要来源于三个方面：①日常工作记录。包括病历、卫生监测记录、健康检查记录等。应注意资料的完整性和准确性。病历是临床研究资料的重要来源，由于没有经过研究的设计环节，可能会产生不完整和不准确的情况。②统计报表。包括工作报表、传染病报表等。报表资料的质量取决于填报人员的认识和责任感，使用时应对数据的准确性做出判断。③专题调查或实验。实验和现场调查一般都经过严格的研究设计过程，但应注意收集资料过程中的质量控制和审核。

3. 整理资料（data sorting） 包括对收集到的资料进行检查和整理的过程。一般采用计算机统计软件包对资料进行核查、汇总与整理分析。在输入计算机前，需要对资料进行编码处理（coding），例如，可以用汉字、字母（如M代表男、F代表女）或数字（如1代表男，2代表女）表示性别。应根据下一步的统计分析选择合适的编码类型。注意输入计算机的信息准确，特别是数量较大、项目较多的资料，可以选择能提高输入质量的数据管理软件如EpiData等软件输入数据。

4. 分析资料（data analysis） 统计分析可以分为统计描述和统计推断两大类。统计描述（statistical description）是对已知的样本（或总体）的分布情况或特征进行分析表述，常用的统计描述方法有统计图（statistical graph）、统计表（statistical table）、统计指标（statistical index）和统计模型（statistical model）等。在统计指标中，常用集中趋势（central tendency）、离散趋势（tendency of dispersion）和相对数（relative number）等表示。

统计推断（statistical inference）是根据已知的样本信息来推断未知的总体，是统计分析的目的，包括参数估计（parameter estimation）和假设检验（hypothesis test）。

【知识点 1-2】

1. 统计工作的基本步骤：研究设计、搜集资料、整理资料和分析资料。

2. 科研结果的好坏取决于研究设计的好坏，研究设计是统计工作的基础和关键，决定着整个统计工作的成败。

3. 统计分析包括统计描述和统计推断。统计描述是对已知的样本（或总体）的分布情况或特征值进行分析表述；统计推断是根据已知的样本信息来推断未知的总体。

第三节 统计资料的类型

【例 1-2】

某医师观察中药溃疡灵治疗成人胃溃疡的疗效，用胃舒平作对照。在进行临床实验设计时，考虑观察病人的年龄（实际岁数）、性别（男、女）、民族（具体民族）、文化程度（文盲、小学、初中、高中、中专、大专、本科、硕士研究生、博士研究生）、职业（无业、个体、农民、工人、干部）、血型（A、B、O、AB）、病情（轻、中、重）、病程（天）、血常规（红细胞计数、血红蛋白、白细胞计数等）、临床治疗效果（治愈、显效、有效、无效）等30个指标。随机抽取成人胃溃疡患者200人作为研究对象，随机分成治疗组和对照组，治疗组用中药溃疡灵治疗100人，治愈40人，显效30人，有效15人，无效5人，总有效95人，总有效率为95.0%；对照组用胃舒平治疗100人，治愈20人，显效15人，有效35人，无效20人，总有效80人，总有效率78.3%，治疗组和对照组的疗效分布差异有统计学意义（$z=-4.876$，$P<0.001$），治疗组治愈率高于对照组；治疗组和对照组的总有效率比较差异有统计学意义（$\chi^2=10.286$，$P=0.001$），治疗组总有效率高于对照组。

聪明在于学习，天才在于积累。

——华罗庚

在上述例 1-2 中，观察对象的指标因个体差异不同而统称为变量（variable），如病人的年龄、性别、职业等，变量可分为定性与定量两种类型，前者说明事物的类别和本质，后者反映事物的数量特征。兼具两种性质者常称为半定量变量或等级变量。变量观察结果或变量的测定数值称为变量值（variable value），如实际的年龄、性别的男女等。某次研究变量值的组合构成了该次研究的统计资料（data）。由不同的变量产生的统计资料类型也不同，由定量变量产生的统计资料一般称为计量资料（measurement data），而由定性变量产生的统计资料一般称为计数资料（count data）或定

性资料（qualitative data）。等级变量产生的资料是等级资料（ordinal data）。

医学领域原始资料类型可分为计量资料（measurement data）、计数资料（count data）和等级资料（ordinal data），三类资料在一定条件下可以互相转换。

不读书的人，思想就会停止。

——狄德罗

1. 计量资料（measurement data，或定量资料 quantitative data） 计量指标也称数值变量（numerical variable）或定量变量（quantitative variable）。计量资料是用定量的方法对每一个观察单位的某项指标进行测定所得的资料。计量资料的变量值是定量的，表现为数值大小，一般具有度量衡单位，例 1-2 中年龄、RBC 数、血红蛋白等。定量变量按取值的不同可分为离散型变量（discrete variable）和连续型变量（continuous variable）两种，前者取值范围是有限个值或者一个数列构成的，常取 0 和正整数值，如现有子女数，儿童的龋齿数，胎次等。连续型变量则可以取实数轴上的任何数值，如身高、体重、血红蛋白等。

2. 计数资料（count data，或定性资料 qualitative data，或分类资料 categorical data） 计数指标也称分类变量（categorical variable）或名义变量（nominative variable）。计数资料是把观察单位按某种属性（性质）或类别进行分组，清点各组观察单位数所得资料。各观察数值是定性的，一般无度量衡单位。各属性之间互不相容。例 1-2 中的性别、职业、血型等

3. 等级资料（ordinal data） 等级资料是把观察单位按属性程度或等级顺序分组，清点各组观察单位数所得资料。各属性之间有程度的差别。各属性之间互不相容。例 1-2 中的文化程度、临床治疗效果等。等级资料的等级顺序不能任意颠倒。两分类（dichotomy）的等级资料由于分析方法与计数资料相同，因此，等级资料通常指有序多分类的资料。

变量和资料的类型以及相应的统计方法列于表 1-1。

表 1-1 变量或资料的类型及其相应的分析方法

变量类型	变量值表现	资料类型	例子	可选分析方法
定量变量（数值变量）				
离散型变量	不连续的数值	计量资料	出生孩子数、死亡动物数等	t 检验、方差分析、相关回归分析等
连续型变量	连续的数值	计量资料	身高、体重、血红蛋白、血清铁含量等	t 检验、方差分析、相关回归分析等
分类变量				
无序分类：二分类	定性（不同属性）对立的两类	计数资料	性别	χ^2 检验、z 检验等
多分类	类间无程度差异	计数资料	血型、职业	χ^2 检验等
有序分类	类间有程度差异	等级资料	文化程度、临床治疗结果	秩和检验、Ridit 分析等

【知识点 1-3】 **医学原始资料类型**

1. 计量资料是用定量的方法对每一个观察单位的某项指标进行测定所得的资料。
2. 计数资料是把观察单位按某种属性（性质）或类别进行分组，清点各组观察单位数所得资料。
3. 等级资料是把观察单位按属性程度或等级顺序分组，清点各组观察单位数所得资料。各属性之间有程度的差别。等级资料的等级顺序不能任意颠倒。

第四节 统计学的几个基本概念

1. 随机事件与必然事件 随机事件（random event）是指随机现象的某个可能观察结果或可能发生也可能不发生的事件，如医疗事故、交通事故等。必然事件（certain event）是指一定要发生的事件，如水加温到 100℃就成开水了。

2. 同质与变异 统计分析是建立在同质基础上的。同质（homogeneity）是指所研究的观察对象具有某

些相同的性质或特征。变异（variation）是同质个体的某项指标之间的差异，即个体变异（individual variation）或个体差异性。如研究儿童的生长发育情况，研究对象是同年龄、同性别的儿童，其生长发育指标如身高、体重、智商等各自不同，即存在个体变异；研究某种新药对高血压病人的疗效，研究对象为确诊的高血压病人，使用相同的药物进行治疗，其疗效也不尽相同。统计分析的目的就是在同质的基础上对变异进行研究，找出客观存在的规律性，从而对同类事物加以估计和预测，以便指导实际工作。

学无早晚，但恐始勤终随。
——［宋］张孝祥

3. 总体与样本 总体（population）是根据研究目的确定的同质研究对象的全体（或全部同质观察单位）。观察单位数有限的总体称为有限总体（finite population），如某校2016年在校大学生的体质调查，某地2016年高血压流行病学调查等。无法确定数量的总体称为无限总体（infinite population）。

要研究总体中的全部观察单位，需要花费巨大的资源，有时是不可能的。在实际工作中，常常是从总体中抽取一部分有代表性的个体组成样本（sample），对样本进行研究以推断总体。样本是从总体中抽取的具有代表性的部分个体（individual），其能否代表总体取决于抽取样本的过程，即抽样（sampling）。样本的代表性还取决于抽取的个体数的多少，称为样本含量或样本例数（sample size）。抽样方法和样本含量估计详见调查设计。

总体与样本的划分是相对的，一个研究中的样本可能是另一个研究中的总体。

4. 抽样研究与抽样误差 通过从总体中随机抽取样本，对样本信息进行分析，从而推断总体特征的研究方法称为抽样研究（sampling research）。由随机抽样造成的样本指标与总体指标之间、样本指标与样本指标之间的差异称为抽样误差（sampling error）。抽样误差的根源在于个体变异，在抽样研究中是不可避免的，但其规律可以认识。统计学的任务之一就是寻找抽样误差的规律并估计其大小。

5. 参数与统计量 反映总体特征的指标称为参数（parameter），常用小写的希腊字母表示，确定的研究总体的参数是常数。而通过样本资料计算出来的相应指标称为统计量（statistic），常用英文字母表示。根据统计量推断参数是统计推断（statistical inference）的主要任务，包括对总体参数大小进行估计的参数估计（parameter estimation）和对总体参数进行比较的假设检验（hypothesis test）。

书山有路勤为径，学海无涯苦作舟。
——［唐］韩愈

6. 概率 概率（probability，P）是随机事件发生可能性大小的数值度量。概率的取值为$0 \leqslant P \leqslant 1$。$P=1$的事件称为必然事件，即一定会发生的事件；而$P=0$的事件为不可能事件，即不会发生的事件；$P$介于0～1之间的事件称为随机事件，即可能发生也可能不发生的事件。$P \leqslant 0.05$的随机事件称为小概率事件，小概率事件的原理是在一次实验中是不大可能发生的。统计学研究的事件只是随机事件，必然事件和不可能事件不属于统计学的研究范畴。

【知识点1-4】 **统计基本概念**

1. 总体是根据研究目的确定的同质研究对象的全体。样本是总体中具有代表性的一部分个体。

2. 抽样研究是通过从总体中随机抽取样本，对样本信息进行分析，从而推断总体的研究方法。抽样误差是由随机抽样造成的样本指标与总体指标之间、样本指标与样本指标之间的差异。其根源在于总体中的个体存在变异性。只要是抽样研究，就一定存在抽样误差，不能用样本的指标直接下结论。

3. 统计学的主要任务是进行统计推断，包括参数估计和假设检验。

4. 概率是某随机事件发生可能性大小（或机会大小）的数值度量。小概率事件是指$P \leqslant 0.05$的随机事件。

思考练习

一、名词解释

1. 总体与样本　　2. 参数与统计量　　3. 抽样研究与抽样误差　　4. 概率

二、是非题（正确记“+”，错误记“-”）

1. 只要增加样本例数就可以避免抽样误差。（　　）

2. 某医院发生的医疗事故属于小概率事件。（ ）
3. 统计描述就是用样本推断总体的统计过程。（ ）
4. 如果对全部研究对象都进行了调查或测定就没有抽样误差。（ ）
5. 分类资料中的各类别可以相互包含。（ ）
6. 医学领域中的三类资料不能互相转换。（ ）
7. 定量变量按取值的不同可分为离散型变量和连续型变量两种。（ ）
8. 科研结果的好坏取决于研究设计的好坏，研究设计是统计工作的基础和关键，决定着整个统计工作的成败。（ ）
9. 没有较好的统计学知识，就不可能进行较好的科学研究，更不可能写出一篇高质量的科研论文。（ ）
10. 用 SAS 和 SPSS 统计分析的结果，在国际学术交流中可以不必说明算法。（ ）

三、选择题（从 a～e 中选出一个最佳答案）

1. 若成年男子以血红蛋白＜125g/L 为贫血，调查某地 1000 人中有多少个贫血患者，这是_____。
a. 计量资料　b. 还不能决定是计量资料还是计数资料　c. 计数资料
d. 既可作计量也可作计数资料　e. 等级资料

2.一批病人的淋巴细胞转换率（%）是_____。
a. 计量资料　b. 还不能决定是计量资料还是计数资料　c. 计数资料
d. 既可作计量也可作计数资料　e. 等级资料

3.统计一批糖尿病患者的住院天数是_____。
a. 计量资料　b. 还不能决定是计量资料还是计数资料　c. 计数资料
d. 既可作计量也可作计数资料　e. 等级资料

4. 测量某病患者的抗体滴度（1∶2，1∶4，1∶8，…），是_____。
a. 计量资料　b. 还不能决定是计量资料还是计数资料　c. 计数资料
d. 既可作计量也可作计数资料　e. 等级资料

5. 调查某医院医生的工作状况，医生一天内上班的时间是_____。
a. 变量　b. 总体　c. 个体　d. 变量值　e. 统计指标

6.治疗结果分为有效和无效的资料，严格说来属于_____。
a. 等级资料　b. 计数资料　c. 计量资料　d. 等级或计量均可　e. 计数或计量均可

四、简答题

1. 某医师根据自己 20 年来收集的胆结石病例进行分析，认为：胆结石的发病和居住地有关，某些地区特别容易发生胆结石。女性发生胆结石的机会比男性大。从治疗效果看，保守治疗的效果不如手术治疗的效果好。

请从统计学的角度，分析该医师的结论。

2. 举例说明如何正确区分不同类型的统计资料？

3. 举例说明如何进行不同类型资料间的相互转换？

（罗家洪　彭林珍　罗　健）

第 2 章　计量资料的统计描述

【例 2-1】

某内科医生调查得到 100 名 40～50 岁健康男子总胆固醇（mg/dl），结果如下：

138	149	155	156	161	163	167	167	171	172
172	172	174	174	174	175	178	180	181	181
184	185	186	186	186	189	189	190	190	190
193	193	194	195	195	196	197	197	197	198
199	199	199	199	199	200	201	202	202	203
203	203	206	207	207	208	208	209	209	209
210	210	213	214	214	216	217	220	220	222
224	224	225	226	227	230	231	232	234	234
235	235	235	236	238	244	246	246	247	248
249	253	255	257	259	259	266	273	277	**278**

【问题 2-1】

（1）例 2-1 是什么资料？

（2）这些人的总胆固醇有什么特征？

（3）如何描述这些人的总胆固醇？

> 你想要别人怎样待你，就得先怎样待别人。
> ——［美］戴尔·卡耐基

【分析】

（1）总胆固醇是通过测量得到的具体数字，有度量衡单位，属于计量资料。

（2）由调查或试验收集来的原始资料，往往是零乱的，需要对资料进行整理，首先要考虑如何表达资料，即对资料作统计描述。资料的统计描述一般用统计表、统计图及统计指标。不同类型的资料有不同的描述方法。

（3）对总胆固醇这样的连续型计量资料，我们首先可以采用频数表来进行描述。统计图可选择频数分布图（直方图）。而统计指标的选择将在后面加以介绍。

第一节　频数表和直方图

一、频　数　表

表 2-1　某市 130 名初中女生一分钟仰卧起坐完成次数频率分布

次数	人数	频率（%）	累计频率（%）
15	6	4.62	4.62
16	9	6.92	11.54
17	18	13.85	25.38
18	28	21.54	46.92
19	33	25.38	72.31
20	25	19.23	91.54
21	8	6.15	97.69
22	3	2.31	100.00
合计	130	100.00	—

用于描述计量资料的分组组段及其频数（frequency）的统计表称为频数分布表，简称频数表（frequency table）。对于离散型资料，只要列出取值及其相应的例数（即频数），就完成了频数表的编制，也可计算相应的频率（%）及累计频率（%）（表 2-1）。

对于类似总胆固醇的连续型资料，以例 2-1 为例进行频数表的编制。

（一）求极差

极差（range）是资料中的最大值（maximum value）与最小值（minimum value）之差，又称为全距，用 R 表示。

$$R = \max(x) - \min(x) \qquad (2\text{-}1)$$

本例中最大值为 278 mg/dl，最小值为 138 mg/dl，故极

差 $R=278-138=140$（mg/dl）。也就是说这 100 个人的总胆固醇值最大相差 140mg/dl。

（二）确定组数

组数的多少视样本含量及资料的变动范围大小而定，一般以达到既简化资料又不影响反映资料的规律性为原则。组数要适当，不宜过多，亦不宜过少。分组越多所求得的统计量越精确，但增大了运算量；若分组过少，资料的规律性就反映不出来，计算出的统计量的精确性也较差。通常分成 8～15 个组（一般样本量在 100 左右的分 10 组，样本量较大时，组数可适当增加）。本例拟分 10 组。

（三）确定组距

每组上限（upper limit）与下限（lower limit）之差称为组距（class interval）。组距可以相等，也可以不等，实际应用时一般采用等距分组。组距的大小由极差与组数确定。

$$\text{组距} = \text{极差}/\text{组数} \qquad (2\text{-}2)$$

分 10 组时，用极差的 1/10 并取整数作组距。本例组距=140/10=14，为便于计算，组距可适当取整，本例取整数为 15。

（四）确定各组段的上下限

各组段的起点和终点分别称为该组的下限（lower limit）和上限（upper limit），显然上限=下限+组距。第一组段必须包含最小值，其下限≤最小值。本例最小值为 138，可取 130 为第一组段的下限，因此第一组段的上限=130+15=145。需要注意，各组段不能重叠，第一组段为[130，145），在频数分布表中用 130～表示（见表 2-2），若总胆固醇值恰为 145mg/dl，则应归入下一组段。依此类推。最末一组必须包括最大值，其上限≥最大值。如本例最大总胆固醇值为 278mg/dl，最末组段为 265～280。

（五）归组计数，作频数分布表

分组结束后，按照“下限≤x＜上限”的原则统计各组段内的观测值个数即频数，将各组段及其落入该组段的频数列成相应的频数表（见表 2-2 第 1、2 列）。在编制频数表时，可计算相应的频率（表 2-2 第 3 列）、累计频率（表 2-2 第 5 列）等指标。

表 2-2 某地 100 名 40～50 岁健康男子总胆固醇（mg/dl）值的频数分布

组段	频数	频率（%）	累计频数	累计频率（%）
（1）	（2）	（3）	（4）	（5）
130～	1	1.0	1.0	1.0
145～	3	3.0	3.0	4.0
160～	11	11.0	11.0	15.0
175～	12	12.0	12.0	27.0
190～	25	25.0	25.0	52.0
205～	15	15.0	15.0	67.0
220～	13	13.0	13.0	80.0
235～	11	11.0	11.0	91.0
250～	5	5.0	5.0	96.0
265～280	4	4.0	4.0	100.0
合计	100	100.00	—	—

据表 2-2，我们可以看出 100 人的总胆固醇值分布在 130～280 mg/dl 之间，以中间 190～204 mg/dl 范围内人数最多，占 25.0%。而总胆固醇值较低和较高的人数都逐渐减少。

二、直 方 图

除了频数表，计量资料还可以用频数分布图来描述。对表 2-1 的离散型计量资料，在横轴上取变量值，纵轴表示相应的频数，用等宽的长方形表示各变量值的频数，由于变量值是不连续的，因此，各长方形间有一间隔，这种统计图称为条图（bar graph），如图 2-1。

对连续型变量，以横轴表示变量值，纵轴表示频数，也用等宽的长方形代表各组的频数，但变量值是连续的，因此长方形也是相连的，这种频数分布图，又称直方图（histogram）。例 2-1 资料的直方图如图 2-2。频数分布图的用途与频数表类似，但图形更直观和形象。

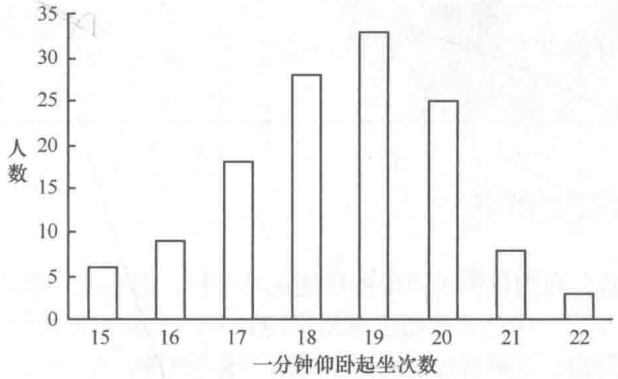


图 2-1　某市 130 名初中女生一分钟仰卧起坐完成次数

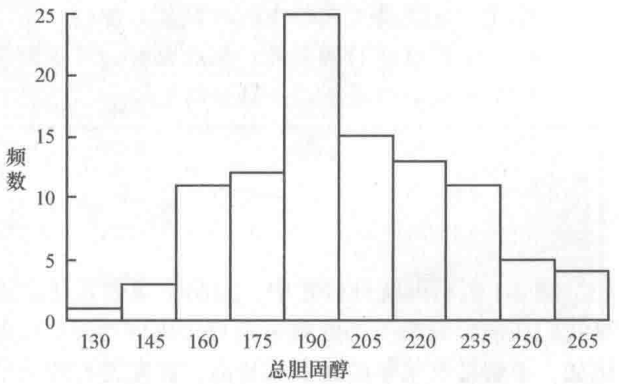


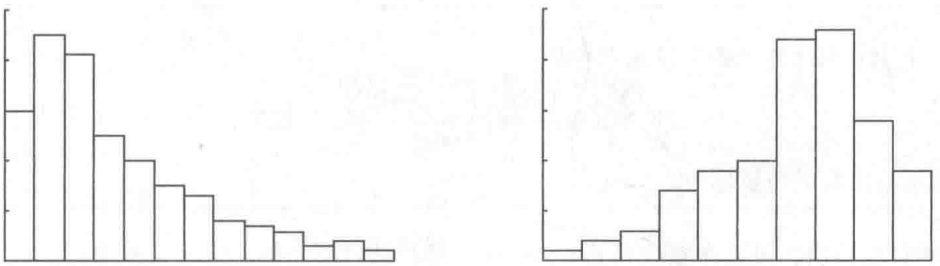
图 2-2　参加慢性病调查的 100 人所测总胆固醇值的频数分布图

> 古之立大事者，不惟有超世之才，亦必有坚忍不拔之志。
>
> ——［宋］苏轼

从直方图我们可以看出，100 人的总胆固醇值的频数在 190～组段较集中，该组段的频数最大，称为集中趋势（central tendency），以此为中心，两侧（收缩压较高与较低）的频数逐渐减少，这种同质的一组数据的分散程度称为离散趋势（dispersion）。

频数表和频数分布图的用途：

（1）揭示计量资料的分布类型。从图 2-2 可以看出，100 人的总胆固醇分布，频数分布的高峰在中间，两端基本对称，逐步减少，这种分布称为近似正态分布，如果两端完全对称则称为正态分布（normal distribution）。如果高峰偏离中心，称为偏态分布（skew distribution），如果高峰在左侧（小的一侧）称为正偏态（positive skew）分布，如果高峰在右侧（大的一侧）称为负偏态（negative skew）分布（图 2-3）。根据资料的分布类型选择合适的统计描述指标。



A. 正偏态　　B. 负偏态

图 2-3　偏态分布示意图

（2）揭示计量资料的资料分布的重要特征：集中趋势（central tendency）和离散趋势（dispersion）。从频数表和频数分布图可以看出计量资料的两个特征：一方面，所有的观测值以某一数值为中心，即频数分布有一个高峰，称为集中趋势，反映资料的平均水平或中间位置；另一方面，观察值又不同程度地偏离集中位置，即存在离散趋势或资料的变异程度。因此，在描述资料时，要全面描述计量资料，需要对资料的集中趋势和离散趋势都进行描述。

（3）便于发现某些特大或特小的可疑值。如在频数表的两端连续出现几个组段的频数为 0 后，又出现一个特大或特小的值，应该怀疑这个数值在测量上可能有误，提醒研究者进一步检查核实。

（4）作为陈述资料的形式。如医院传染病年统计报表、学生成绩频数分布等。例数大时，可以频率估计概率。

（5）提供分组数据，便于进一步计算统计描述指标和统计分析。

【知识点 2-1】 **频数表和频数分布图的用途**

1. 揭示计量资料的分布类型。
2. 揭示计量资料分布的重要特征——集中趋势与离散趋势。
3. 便于发现特大或特小的可疑值。
4. 作为陈述资料的形式。例数大时，可以频率估计概率。
5. 便于资料的进一步统计分析。

第二节 集中趋势的描述

例 2-1 资料的统计分析中，编制了频数表及绘制频数分布图仅做了简单的描述统计分析，需要进一步计算我们在研究时最关心的集中趋势和离散趋势指标的大小，集中趋势的指标常用平均数（average）来表示。例如，了解某地某年龄儿童的身高，首先关心的是平均身高；了解某癌症患者手术后的存活时间，首先关心术后平均存活时间等。

平均数（average）是描述一组同质变量值的平均水平或集中趋势的指标，也可描述频数分布的集中位置。常用的平均数有算术均数（arithmetic mean）、几何均数（geometric mean）、中位数（median）、众数（mode）和调和均数（harmonic mean）。本教材只介绍前三个平均数。

一、算术均数

算术均数（arithmetic mean）简称均数（mean），即所有观察值的和除以观察值的个数。总体均数用 μ 表示，描述总体的集中趋势；样本均数用 $\bar{x}$（念作 x bar）表示，描述样本资料的集中趋势。

设某一资料包含 n 个观测值：x_1，x_2，…，x_n，则样本均数 $\bar{x}$ 为：

$$\bar{x}=\frac{x_1+x_2+\cdots+x_n}{n}=\frac{\sum_{i=1}^{n}x_i}{n}=\frac{\sum x}{n} \qquad (2\text{-}3)$$

【例 2-2】

求例 2-1 中 100 人的总胆固醇值算术均数。

$$\bar{x}=\frac{\sum x}{n}=\frac{138+149+\cdots+278}{100}=207.41$$

100 人的总胆固醇平均值约为 207.41mg/dl。

对于分组的计量资料，可以在频数表的基础上采用加权法计算均数，计算公式为：

$$\bar{x}=\frac{f_1x_1+f_2x_2+\cdots+f_kx_k}{f_1+f_2+\cdots+f_k}=\frac{\sum_{i=1}^{k}f_ix_i}{\sum_{i=1}^{k}f_i}=\frac{\sum fx}{\sum f} \qquad (2\text{-}4)$$

式中，x_i 为第 i 组组中值，$x_i=\dfrac{\text{组段下限}+\text{组段上限}}{2}$；$f_i$ 为第 i 组频数；k 为组数。

第 i 组的频数 f_i 是权衡第 i 组组中值 x_i 在计算均数时所占比重的大小，因此 f_i 称为 x_i 的“权数”，加权法由此得名。