高等医药院校协编教材

供硕士学位研究生、七年制学员用

医学统计学

蒋知俭　主编

人民卫生出版社

高等医药院校协编教材

供硕士学位研究生、七年制学员用

医学统计学

蒋知俭　主编

人民卫生出版社

图书在版编目(CIP)数据

医学统计学/蒋知俭主编. —北京：人民卫生出版社，1997

ISBN 7-117-02704-5

Ⅰ.医… Ⅱ.蒋… Ⅲ.医学统计-统计学 Ⅳ.R195.1

中国版本图书馆 CIP 数据核字（97）第 09311 号

医 学 统 计 学

蒋 知 俭 主编

人民卫生出版社出版发行

（100078 北京市丰台区方庄芳群园 3 区 3 号楼）

北 京 人 卫 印 刷 厂 印 刷

新 华 书 店 经 销

787×1092 16开本 37$\frac{3}{4}$印张 877千字

1997年8月第1版 1998年6月第1版第2次印刷

印数：5 071—6 570

ISBN 7-117-02704-5/R·2705 定价：40.00元

（凡属质量问题请与本社发行部联系退换）

编写委员会

主　　编　蒋知俭

副 主 编　王润华　刘桂芬　李　康　谭　瀛
刘　钢　李颖琰　孙晓光

编写人员（按姓氏笔画排列）

王全丽　王润华　史元峰　田考聪　刘言训　刘　钢
刘桂芬　孙晓光　李炽民　李　康　李颖琰　何大卫
庞维秋　周燕荣　赵欣涛　姜　虹　冯丽云　曾　庆
蒋知俭　谭　瀛

编 写 说 明

研究生教育是我国教育结构中最高层次，为提高学位研究生的培养质量，教材建设是十分重要的环节。目前国内尚无统编及系统教材。鉴于此，山西医科大学、山东医科大学、白求恩医科大学、西安医科大学、河南医科大学、重庆医科大学、哈尔滨医科大学等七院校在互通信息、充分酝酿，并对主要观点取得一致看法的基础上，成立《医学统计学》协编教材委员会，着手编写供硕士研究生（含七年制）使用的教材。

本书的编写主要从实际出发，在医学院校本科所学课程基础上，侧重思维方法和分析能力的培养。全书共32章，按130学时编排。主要内容包括：基本统计方法（30学时）、医学实验设计（30学时）、医用多因素分析（20学时）、电子计算机软件（50学时）等四部分。各章后附有习题，有*号的节可供选修，书后附有工具表、英汉医学统计词汇及参考文献，供学员在今后工作中有较高的参考价值。

随着医学模式的转变和医学教育改革的深化，我们积多年实践经验，参考国内外部分研究生培养教学大纲，认为本书应主要立足于培养目标，强调理论联系实际，在内容上既要反映本学科的最新进展，更应注重基本理论，基本知识和基本技能的培养。本书的编写以概念清楚、方法准确、重在应用为主要特色。

在编写过程中，编委会两次集会，充分交流了编写方案，最后通过集中讨论审稿和定稿。我们希望《医学统计学》教材能达到预期目的。

由于我们的学识和水平有限，难免有缺点和错误，恳请读者及使用单位批评指正。

蒋知俭

1996年9月于哈尔滨

目　录

第一篇　基本统计学方法

△ 第二篇 调查设计和实验研究设计

第三篇　多元统计分析

第四篇 SAS 统计软件应用

第一篇　基本统计学方法

第一章　绪　　论

第一节　医学统计学的作用和内容

随着医学的发展，医学统计学已逐渐为广大医务工作者所认识，并广为应用，成为医学科学研究的重要手段。无论是基础医学、临床医学和预防医学的各个领域的科学研究，还是疾病防治计划的拟定、效果评价、调查与实验研究的设计，医学统计学都为资料的收集，整理与分析以及疾病预测等提供了有效的工具。在现场调查研究、实验研究和临床试验中都广泛采用着现代统计学方法。电子计算机的应用，促进了多变量分析等统计方法在医学研究中的开发和利用。

医学统计学是运用概率论和数理统计的原理、方法紧密结合医学实践，研究医药卫生领域中资料的收集、整理、分析和推断的一门应用学科。

医学统计学的主要内容包括：

1. 调查与实验研究设计　科研设计的质量直接影响着实验结果的准确性、可靠性、严密性和代表性，是实验数据处理的前提，决定着科学研究的成败。一个完整的科研设计包括专业设计与统计设计。专业设计是指研究者对专业知识的把握能力，直接影响着实验的深度和水平；统计设计是指研究者对统计知识的正确应用，直接影响着科学实验的质量，两者相互结合，缺一不可（表 1-1）。

表 1-1　调查与实验研究设计

	专　业　设　计	统　计　设　计
要求	运用专业知识进行设计	运用统计学知识进行设计
内容	选题、调查（实验）、方法、材料设备、环境及指标选择等	确定设计方案、收集整理资料、确定统计指标、分析与推断方法等
方向	探讨实验、观察结果的适用性和创造性	探讨实验、观察结果的可重复性、高效性
目的	回答和解决科研课题、验证假说，保证科研成果的先进性	减少和控制误差，保证样本的代表性和可靠性；保证实验结果的精确性和可重复性

调查与实验设计主要包括：①立题；②制定研究计划；③明确设计的三个要素（研究对象、处理因素、实验效应）；④保证研究成功的条件（设对照、控制误差、确定样本含量及抽样方法、确定研究方式、设计方案、保证手段、预期后果及分析方案等）。主要

设计方案有：自身对照设计、配对设计、完全随机设计、随机区组设计、析因设计、拉丁方设计、正交设计、序贯设计及多元分析设计等；此外，还包括某些专项设计（如现场调查设计、临床实验设计、动物实验设计等）。

2. 医学统计学方法　运用统计学的原理和方法研究医学领域中的生物、理化、社会、心理等因素及机体的内、外环境条件对人体健康的影响，认识人群健康和疾病现象的数量特征。主要内容包括：①统计资料的收集与整理；②常用统计指标：集中趋势（算术均数$\bar{X}$、几何均数G、中位数M、众数等）与离散趋势（标准差S、方差S^2、变异系数CV、极差R、四分位数间距Q、平均差A等）；相对数（率、构成比、动态数列）；相关系数r、回归系数b；半数效量ED_{50}、半数致死量LD_{50}；相对危险性RR以及绝对数等；③分析资料：计算标准误进行参数估计，据资料的性质选择检验方式（t检验、u检验、χ^2检验、F检验、非参数检验、Ridit分析等）；④统计图与统计表。

3. 医学多元统计方法　医学现象复杂多变，如疾病的发生、病情的变化、转归、预后等往往包含着众多因素的作用，为充分运用观察资料的综合信息、分析其因果关系、内在联系的统计规律，作出科学的符合实际的结论采用多因素分析的方法。主要内容包括：多元线性回归、逐步回归、判别分析、聚类分析、主成分分析、因子分析、典型相关分析、logistic与Cox回归分析等。

4. 计算机统计软件　近30年来，电子计算机技术发展迅速，应用日趋广泛，已成为医学统计学的重要手段。医学是研究生命现象和疾病防治的规律，涉及面广、技术手段多样化、信息量大。电子计算机的应用为大量的信息储存与检索、复杂数据的处理、抽样模拟等，尤其是对多因素分析的开展提供了条件。不少多元分析的计算程序相继问世，形成软件包。本书主要介绍SAS统计软件，并和前述内容同步。主要包括：SAS统计软件概述、SAS数据集、基本统计分析、多元统计分析等。

第二节　统计资料的收集与整理

统计工作一般分为三个阶段：收集资料、整理资料和分析资料。三者密切联系，不可偏废。

1. 统计资料的来源　收集原始资料是统计工作最重要的一步，关系着整个统计工作的质量。医学科学研究主要通过调查或实验来收集资料，要求及时、准确、完整。医学统计资料的来源，主要有：统计报表、医疗工作的原始记录和报告卡、专题调查或实验等三个方面。

（1）统计报表：医疗卫生工作报表是用表格形式，按一定时间和程序，系统地收集统计资料的组织形式。如医院工作年报表，传染病、疫情、居民出生、病伤死亡报表等是依据国家规定的报表制度由医疗卫生机构定期逐级上报。要求填写认真、准确、及时上报，防止漏报错报。统计报表是了解居民健康状况的基础资料，为拟订医疗卫生工作计划与措施提供科学的依据。

（2）日常医疗卫生工作记录：医疗卫生工作的原始记录是业务管理和科学研究的重要资料，是一种经常性工作记录，如门诊病历、住院病历、健康检查记录、医学检查记录等，要求完整、准确、认真填写。医院病历是诊治疾病的工作记录，是研究疾病病情和疗效的基础资料，使用时应根据设计要求进行挑选，并注意病例往往因时间与空间不

同而异，有时有一定的局限性。疾病分类是依据防治疾病任务的需要，按病因、解剖部位等特点进行归纳和概括的，我国现行的疾病和死因分类是卫生部1987年在国际疾病分类ICD-9基础上结合我国具体条件而修订的。

（3）专题或实验研究资料：是根据业务和科研的需要，专门选用某些调查或实验方法，有明确针对性地收集资料，是医学研究中常用的方法。通常依据研究目的确定研究对象、指标、研究方式、组织计划、时间与地域分布，确定调查表格（含一般项目、研究项目和研究者的项目等）。如恶性肿瘤、心脑血管疾病的流行病学调查，某种药物疗效、手术疗法的临床观察等资料的收集。统计调查可分为常规性和专题性调查两种类型（表1-2）。

表1-2　医学统计调查基本过程表解

以说明人民健康水平与特征为目的的常规性调查	以探讨事物的新内容、新领域或新联系为目的的专题性调查
↓	↓
根据统计报表、医疗卫生工作的原始记录等收集现有的有关资料整理确定： ①与致病因子、环境条件及人类自身情况有关问题 ②关于疾病发生与扩延情况	根据历史文献、有关统计数字、口头流传资料收集可信资料分类确定： ①问题之所在 ②评价现有资料 ③构成科学假设 ④通过实际调查、验证假设
↓	↓
①计算反映人民健康水平的一系列指标 ②说明健康与疾病的动态 ③对人民健康水平与疾病扩延程度作出诊断	①从人体内外环境因素研究疾病发生与扩延的原因 ②评价防治措施效果 ③制定防治措施，并考虑付诸实现的形式
↓	↓
提出问题或尚待解决的问题	解决问题
↘	↙
开展更有效的医疗卫生措施，发展医学理论	

单或双因素设计的原始资料的收集要保证除研究外其他因素（干扰因素）齐同条件，故收集有一定困难，而多因素设计原始资料的收集可把众多因素纳入研究因素内，故可不考虑难于控制的干扰因素，在分析时加以处理。如冠心病病因与多种因素有关，可经多因素分析加以筛选、剔除，找出主要因素。

2. 统计资料的整理　统计资料的整理是将收集来的原始资料进行科学的分组归纳，使资料系统化、条理化，便于进一步计算统计指标和进行统计分析，以反映被研究事物的规律性。整理资料包括审核资料、设计分组、拟整理表和归纳汇总四个步骤。

（1）审核资料：主要是核对原始资料的准确性、完整性和可靠性。方法有两种：一是逻辑检查，依据专业知识和从资料的相互关系中检查是否合乎逻辑、自相矛盾；一是计算检查，计算各横、纵行的合计是否有误，还可和经验数字、参考值比较，找出问题所在。

不合理或错误的项目必须复查、补正或舍弃。在专题调查时，检查资料应在现场调查时进行，边调查边核实，以便用时改正。在实验工作中，要随时核对记录的正确性、完整性。

（2）设计分组：将性质相同的资料归纳到一起，使资料系统化，以反映事物的本质。

分组是统计工作的基础，只有坚持“同质”的原则，才能得出正确的结论，分组可分为质量分组和数量分组两种形式。

1）质量分组：按事物的性质、特征或类型分组，如疾病分类、死因分类；病人按性别、职业、工种分类；疾病的病情按轻、中、重分组；疗效按治愈、显效、好转、无效、恶化分组；化验检查按阳性、阴性或－、±、＋、＋＋、＋＋＋分组等，其特点是同质，组间界限分明。

2）数量分组：在质量分组的基础上，再按变量值的大小来分组。如年龄、身高、体重、血压、心率等，特点是从量的变化分析事物的差别和规律。数量分组的粗细及组数的多少以能说明资料的规律性为准，为便于资料间相互比较，还必须注意到习惯分组法，如年龄分组由0岁开始，习惯按每5岁或10岁一组，身高多间隔2cm一组等。数量分组界限要清楚，除第一组外一般采用下限分组法，即以各组段的下限表示，如用0－，1－，5－，10－，……等表示，由零开始到不满1者归入“0－”组，满1到不足5者归入第二组，余类推。

3. 拟整理表　本表是一种过渡性表格，可表达资料的分配情况和内部结构，初步显示各项目之间的联系，整理项目中要有“其他”或“不详”，“合计”或“总计”等项，以便互相核对（表1-3）。

表1-3　流脑各种病型与病情的关系（整理表）

病　型	病　情			合　计
	轻	中	重	
菌血型				
脑膜炎型				
混合型				
不　详				
合　计				

4. 整理资料　亦称统计归纳，是依据研究设计中整理与分析计划的要求进行分组与汇总，一般常用于手工归纳。手工归纳又分为划记法和分卡法：划记法是将调查表中同类资料逐个记入整理表中，记数的方式常用“正”字或“＋＋＋＋”划记。其优点是经济方便，适用于少量资料，缺点是容易划错。分卡法是将原始调查卡片按分组项目分别归组或过录在一张卡片上再行归组，后清点每组卡片数，填入整理表。资料整理好后，可进一步列出统计表，制统计图，计算统计指标和进行统计分析，作出研究结论。近年来随着电子计算机技术在医学中的应用，大量的医学统计资料、复杂信息的原始数据、病史资料等可进行编码输入电子计算机进行贮存、绘制统计图、计算指标及统计分析，解决复杂的运算。

第三节　统计学中的几个基本概念

一、变量

研究对象中各观察单位个体间的差异称为变异。医学、生物学现象中变异广泛存在，

且十分重要。对某项变异特征进行测量和观察，得到的指标称变量（variable）。如在相同条件下生长的同性别、同年龄的儿童，他们的身高、体重不尽相同，生长速度各异；在病种、病程、病期、病型等方面经严格精选的相同的病人，观察某药物或疗法的疗效，未必相同。变量的测量和观察结果可以是定量的，也可以是定性的，或半定量的，通称观察值或变量值。按变量值的性质可分为不同类型，其统计分析方法也不相同。

1. 数值变量（定量变量） 每一同性质观察单位所具有的变量值是定量的，表现为数值大小，有度量衡单位。如体检时测量的身高、体重、胸围、肺活量、心率、血压及病人的年龄、体温、白细胞数、住院天数等。每个观察单位的观察值之间有量的区别，而同一批观察单位必须是同质的。这种由定量因素（指标）组成的统计资料也称计量资料，它是一群单变量值、双变量值或多变量值。多数数值变量属连续变量。

2. 分类变量（定性变量） 每一同性质观察单位所具有的变量值是定性的，表现为互不相容的类别或属性，又可分为：

（1）无序分类：编制分类资料频数表，清点观察单位数。按分类项目多寡又可分为两项分类和多项分类。两项分类如检查小学生粪便蛔虫卵，计算阳性、阴性数；查清某年某地男、女人口数；感染或未感染某种疾病数。多项分类如测定某人群血型，清点O、A、B、AB型各多少人；分类计数几种主要心血管系统疾病（肺心病、冠心病、高血压性心脏病、风湿性心脏病，先天性心脏病等）人数。分类变量在同组各观察单位之间通常没有量的差别，但组间有质的不同。这种由定性因素（指标）组成的统计资料也称计数资料，属间断性变量。

（2）有序分类：将观察单位按某种属性或某个标志的不同程度，等级分组计数。如对一批病人尿蛋白化验结果按－、±、＋、＋＋、＋＋＋分组；对病人的病情分为轻、中、重；治疗结果通常按治愈、显效、好转、无效、恶化、死亡分组；免疫学中抗体滴度等，其特点是各组之间既有等级顺序，又有程度与量的差别，故也称等级资料或半定量资料。

3. 资料的转化与分析 根据分析问题的需要，各类变量可以互相转化。如作为数值变量的舒张期血压，规定≥12kPa为异常，可清点正常或异常的个数为分类变量；也可按血压过高（>12.7kPa）、临界高血压（12.0—12.7kPa）、血压正常（8.0—12.0kPa）、血压偏低（<8.0kPa）等有序分类。有时亦可将分类变量数量化，如将多项治疗结果转化为评分，分别用0、1、2、…等表示，则可按数值变量分析。

不同类型的变量采用不同的统计方法进行分析。数值变量常用平均数、标准差、t检验、方差分析、相关与回归分析等；无序变量常用率、构成比、χ^2检验；有序分类常用率、构成比、秩和检验、参照单位检验等，上述各类型变量还可以采用多变量（因素）分析法。

二、总体与样本

1. 总体（population） 指同质的研究对象中所有观察单位研究指标变量值的集合。如对某地儿童体温参考值进行研究，研究对象是该地正常儿童，观察单位是每个儿童，变量值为体温测量值，该地全体儿童的体温值即构成总体，该总体是建立在某地14岁以下的正常儿童的同质基础上。总体通常限定于特定的时间与空间范围之内，且为有限数量的观察单位，称为有限总体；有时总体是假设的，没有时间和空间限制，观察单位数是

无限的，称为无限总体，如研究碘盐对缺碘性甲状腺疾病的防治效果；某中药方剂对脑囊虫病的疗效观察等，没有固定的时间和空间限制，观察单位数是不确定的，即总体无限称为无限总体。

2. 样本（sample） 医学实践与研究中，要直接研究无限总体通常是不可能的，即使是有限总体，由于人力、物力、时间、条件等限制，要对其中每个观察单位进行研究或观察，有时也是不可能的，也不必要。而只是从总体中随机抽取部分观察单位，其变量实测值构成样本，目的是用样本指标推断总体特征。如用一滴外周血的化验结果，代表一个人的全血成分；用不同流域、不同浓度的水样水质的各项污染指标代表整个江、河的污染指标等，这种推断需要经过严谨的实验设计，以样本的可靠性和代表性为基础。样本的可靠性：主要是使样本中每一观察单位确属同质总体，如临床实验研究中，需要有正确的诊断，对病种、病程、病期、病型等要严格划分；动物实验中，应规定动物种属、品系、窝别、性别、体重及生理状态力求一致，以排除和控制非实验因素的干扰。样本的代表性：使样本能充分反映总体的实际情况，要求抽样遵循随机的原则，目的是使每个观察单位被抽得的机会相等，避免主观取舍及偏性；还要保证足够的样本含量，即保证足够的观察单位个数。医学研究实际中，通常数值变量可少些，但至少≥7 例才有统计学意义，一般要＞30 例（＜30 例称小样本），最好＞100 例；作为分类变量，样本数量要大些，至少＞30 例，一般＞100 例，常见病、多发病最好为几百例，当然，这些是经验要求，关于样本含量还可进行推算，以后章节将详述。

3. 参数 统计学上描述总体变量的特征称为参数。如总体均数、中位数和众数等描述总体的中心位置或集中趋势；总体标准差、极差、四分位数间距等描述总体的离散趋势等。但总体参数常属未知，而需以样本统计量来估计总体参数称为样本指标。如以样本均数（$\bar{X}$）推算总体均数（μ），以样本标准差（s）推算总体标准差（σ）等，值得注意的是，选择统计量作为参数估计量时，通常应选择无偏、有效且一致的估计量，即对总体变量渐近无偏估计量。

三、概率

1. 随机事件 客观世界中，有许多现象人们可以事先预言它们在一定条件下肯定发生，称必然事件。如刚患过乙型肝炎的人，化验乙型肝炎表面抗原一定为阳性；发作期疟疾病人血中一定能检出疟原虫等。另一类现象，人们不可能事先预言它们必然不可能出现，称为不可能事件。如沙眼患者不能导致死亡；感染过结核病的人，结核菌素试验不会是阴性等。还有更多的现象是事先无法预言其结果一定出现或一定不出现的事件，即在一定条件下具有多种可能发生的结果，而究意发生那一个结果事件不能肯定，即为随机事件，或称偶然事件，如医生用同一种疗法治疗某病病人，结果可以有治愈、好转、无效或死亡。对于刚入院的该病某一病人，治疗后究竟发生那种结果是不确定的。这里的每一种可能结果都是一个随机事件。随机不是随便，在大量多次数的观测或试验中，随机事件的统计规律是可以认识的。

2. 概率（probability） 概率是描述随机事件发生的可能性大小的数值，常用 P 表示。如某研究者观察某降压新药的效果，患者用药后可能有：血压下降、血压无变化、血压升高等情况，研究者颇为关心这些未知数值。可将上述可能的结果“血压下降”这个

事件记为 A，则其概率可记为 $P(A)$，或记为 P。随机事件的概率在 0 与 1 之间，即 $0 \leqslant P \leqslant 1$，常用小数或百分数表示。$P$ 越接近 1，表示某事件发生的可能性越大；P 越接近 0，表示某事件发生的可能性越小。严格说，$P=1$，表示事件必然发生；$P=0$，表示事件不可能发生，它们是确定性的，不是随机事件，但可视为随机事件的特例。统计学上的很多结论都是带有概率性的。习惯上将 $P \leqslant 0.05$ 或 $P \leqslant 0.01$ 称为小概率事件，表示某事件发生的可能性很小。

3. 频率与概率的关系　一个随机试验有几种可能结果，出现某种结果的可能性有多大称为频率。如由中小学教师中随机检诊出喉炎患者 300 名，投予中药制剂治疗，有效率为 87%，这是一个频率。显然我们是通过统计抽样从样本计算频率，然后以频率来近似地估计概率。通常频率并不等于概率，且概率不易求得。当样本含量很小时，自然不能正确地估计概率，而只有样本相当大，才能获得概率的可靠估计值，随着样本不断增大，频率也就越接近概率，当样本含量扩大到等于总体时，频率就等于概率。即当试验次数逐渐增多，频率逐渐稳定在一个常数附近，这个常数称为概率，“频率稳定性”揭示了随机现象的规律性。

4. 概率运算法则　“事件 A 和 B 中至少有一个发生”称此事件为 A、B 之和，记作 $A+B$；“事件 A 和 B 同时发生”称此事件为 A、B 之积，记作 AB；如果 A、B 不可能同时发生，即 A、B 互不相容，事件 A 的对立事件记作 $\overline{A}$。

（1）加法法则：对于任意两事件 A 和 B 中，至少有一个发生的概率为 $P(A+B)=P(A)+P(B)-P(AB)$。若 A、B 互不相容，则 $P(A+B)=P(A)+P(B)$。

相互对立事件的概率之和等于 1，即

$$P(A)+P(\overline{A})=1$$

（2）乘法法则：在“事件 A 已发生”的条件下，事件 B 发生的概率称为的条件概率，记作 $P(B|A)$。

对于任意两事件 A 和 B 同时发生的概率为：

$$P(AB)=P(A)\,P(B|A)$$

或 $$P(AB)=P(B)\,P(A|B)$$

若 A（或 B）发生与否并不影响 B（或 A）的概率，即 A 与 B 相互独立，这时 $P(B|A)=P(B)$ 或 $P(A|B)=P(A)$。若事件 A 与 B 相互独立，则事件 AB 的概率等于 A 的概率与 B 的概率之积，即

$$P(AB)=P(A)\cdot P(B)$$

这一法则可推广到有限个相互独立的事件。

四、几种常见的分布

随机事件有各种不同的结果，变量 X 按不同结果而取不同的值，且 X 服从一定的概率分布，这样的变量称为随机变量。随机变量的分布类型常用有离散型与连续型两种，均可用分布函数表达随机变量的概率。离散型分布如二项分布、Poisson 分布、超几何分布等，多用于分类变量；连续型分布如正态分布、Weibull 分布等，多用于数值变量。

（一）二项分布

1. 定义　在 n 次独立实验中，每次有两个对立的结果（如阳性或阴性），其中某种阳