普通高等教育“十五”国家级规划教材

医学统计学

主编　赵耐青

高等教育出版社
HIGHER EDUCATION PRESS

普通高等教育“十五”国家级规划教材

《医学统计学》

主　　编　赵耐青

副 主 编　陈　峰　夏结来

　　颜　虹　孙振球

编　　者　（以姓氏笔画为序）

王乐三　（中南大学）
王　彤　（山西医科大学）
王洪源　（北京大学）
邓　伟　（复旦大学）
田考聪　（重庆医科大学）
孙振球　（中南大学）
毕育学　（西安交通大学）
李晓松　（四川大学）
张文彤　（复旦大学）
陈冠民　（武汉大学）
陈　峰　（南京医科大学）
陈　智　（中国医科大学）
陆　健　（第二军医大学）
贺　佳　（第二军医大学）
赵耐青　（复旦大学）
夏结来　（第四军医大学）
顾海燕　（南通医学院）
凌　莉　（中山大学）
曹素华　（复旦大学）
颜　虹　（西安交通大学）

学术秘书　张文彤(兼)　邓伟(兼)

高等教育出版社

内容摘要

本书为教育部国家级十五规划教材。内容全面丰富，包括了医学统计的基本概念，计量和计数资料的描述性统计与推断性统计方法，研究设计入门和统计方法学新进展。

为帮助读者学习，本书还给出了3个研究方向的实例：实验室研究、社区调查研究和临床试验研究，使读者可以体会一些实际工作中的统计问题、研究步骤和解决方法。

本书在编写过程中参考国外教材，结合计算机模拟，在书中引入计算机辅助教学的内容，并引入"统计软件Stata实现"内容，在光盘中引入本书所有例题的SAS程序和SPSS程序。参与本教材编写的所有编者均为国内著名院校的统计学专家，有丰富的教学和科研经验。本书适合7年制临床医学、5年制的临床医学、预防医学和其他医学专业本科学生用，也可以作为医科专业研究生教材使用。

图书在版编目（CIP）数据

医学统计学 / 赵耐青主编 . —北京：高等教育出版社，2004.3

ISBN 7-04-013920-0

Ⅰ.医... Ⅱ.赵... Ⅲ.医学统计-高等学校-教材 Ⅳ.R195.1

中国版本图书馆CIP数据核字（2004）第001606号

出版发行	高等教育出版社	购书热线	010-64054588
社　　址	北京市西城区德外大街4号	免费咨询	800-810-0598
邮政编码	100011	网　　址	http://www.hep.edu.cn
总　　机	010-82028899		http://www.hep.com.cn
经　　销	新华书店北京发行所		
印　　刷	高等教育出版社印刷厂		
开　　本	850×1168　1/16	版　　次	2004年3月第1版
印　　张	18.75	印　　次	2004年3月第1次印刷
字　　数	470 000	定　　价	34.00元（含光盘）

本书如有缺页、倒页、脱页等质量问题，请到所购图书销售部门联系调换。

序

由于个体差异，加之自然、社会、心理等各方面的影响，医学现象充满变异。为便于透过变异探索规律，医学统计学已成为医学实践和科研必不可少的工具，以及医学院校公认的重要课程。但是，由于统计学与医学的思维方式有所不同，医学生在学习医学统计学时始终感到概念抽象，颇难学以致用。

随着计算技术的普及，各种功能优异、使用方便的统计软件包纷纷问世，使得计算量惊人的统计分析方法不难实现，通过电脑模拟试验形象地揭示统计学基本原理，加深学生的理解也成为可能。目前，将统计教学与统计软件的使用相结合，强调概念、突出应用、淡化计算已经成为大势所趋，已有数本教材朝这一方向努力。本书也属此列，它采用 Stata 这一短小精悍的统计软件作为教学工具，实现了全书所有统计计算及模拟实验，对医学统计教学做了有益的尝试。

本书的另一特点是注重统计实践，通过给出实验室研究、临床研究和调查研究的实例，使读者了解从研究设计至最后资料分析的整个过程，帮助学生较快地将统计知识运用于医学实践。

本书主编的数理统计功底深厚，编者们长期从事医学统计教学与科研，书中融入了各自的经验。出版在即，承蒙盛情相邀，特撰数言，以之为序。

方积乾
中山大学公共卫生学院
医学统计与流行病学系
2003 年 9 月，广州

前　言

本教材是中华人民共和国教育部国家级“十·五”规划教材。根据教育部教育司批准的本教材标书，我们力图与国外教材接轨；结合计算机软件，适当淡化统计计算；注重理论联系实际，借助计算机辅助教学，强化一些基本概念；我们先在一个班级试用本教材，然后根据试用情况和一些专家评阅意见进行修改，尽可能完善本教材的编写。

本教材适用的对象是七年制、五年制的临床医学专业学生和五年制本科预防医学专业学生。如果对书中的内容略作选择，也可用作其他专业学生的医学统计学教材。

Stata 软件是最适合于教学的一个统计软件，该软件功能强大，操作简便，输出结果针对性极强，并且许多功能都是针对医学研究背景设计的。根据我们的实践经验，只要安排 2～4 个学时介绍 Stata 的 4 个窗口和资料输入方法，就可以让学生根据书中的 Stata 简介，用 Stata 软件独立完成本书例题与习题的有关统计计算和作图，只需提供上机条件而无需另行安排上机课。另外，我们编写了《Stata 软件的基本操作和数据分析入门系列讲座》，置于本教材的配套光盘上，帮助读者较轻松地掌握 Stata 软件进行统计分析。有关 Stata 软件的更详细介绍，读者可以参阅陈峰教授编写的《现代医学统计方法与 Stata 应用》第二版。

本教材配套光盘中提供了一些计算机辅助教学的模拟程序以及其他功能的统计程序。这些程序需要复制到 Stata 目录下的 ADO \ BASE 子目录下，并且在 Stata 命令窗口下输入连接命令 net set ado [d:] \ stata \ ado \ base，其中[d:]表示 Stata 软件所在的路径。

本教材配套光盘中的有关 SAS 内容由罗剑峰老师负责编写，SPSS 内容由张文彤和罗剑峰老师共同负责编写，本教材中的所有图表由董伟老师制作。

本教材的编写得到高等教育出版社、复旦大学各级领导、尤其是校教务处等有关部门的大力支持。我的前辈老师中山大学的方积乾教授、复旦大学曹素华教授和詹绍康教授始终关注这本教材的编写，他们的见解和建议令我受益匪浅，确保本教材更加完善。特别，曹素华教授不仅在编写安排中提出了许多指导性的意见，而且亲自参加编写和审阅，为本教材出版付出了许多心血和智慧。

在此，我衷心地感谢所有参编专家和老师们为本教材所贡献的智慧和经验，以及他们所付出的辛勤劳动。我们真诚合作的友谊必将天长地久。

我特别要感谢陈峰教授为本教材的编写、审阅和定稿所做出的特殊贡献，感谢夏结来教授、陈冠民教授、王彤教授和张文彤老师为本教材编写和审阅所做出的一切努力，感谢研究生张娴静为本教材校对所付出的辛勤劳动。

由于作者能力所限，教材中不免存在不足之处，敬请广大师生提出宝贵意见。

赵耐青

2003 年 9 月于上海

目　录

第一章 绪 论

第一节 统计学与医学统计学

统计学(Statistics)是研究数据苏集、整理、分析、推断等原理和方法的学科。社会生产活动及日常生活中的一些数据汇总和报表只是统计工作中极少的一部分内容。统计学研究内容主要分为两部分:① 参与随机现象研究的设计、观察和资料的收集,并处理一些与统计学相关的问题或提出建议;② 根据数理统计学原理对收集的资料进行统计分析并做出统计推断。

医学科学研究中的许多观察结果都是不能事先确定的,即使条件完全相同的两次观察结果往往也是不同的,所以观察结果具有随机性。因此在医学研究领域中的研究设计、资料收集和结果分析常常需要运用统计学知识。医学统计学(Medical Statistics)就是统计学原理和方法在医学研究领域的应用。20 世纪中期以后,医学统计学逐渐形成一门学科,其在医学研究中的作用也愈显重要。目前,许多国际性医学研究项目均需医学统计学人员参加。我国的《药品注册管理办法》规定新药临床试验必须自始至终有统计学人员参与。目前医学统计学已经成为医学研究领域中的重要组成部分,并是医学各专业本科生和研究生的必修课程。

国际统计学界通常把生命科学研究、临床医学研究和预防医学研究中的统计学内容统称为生物统计学(Biostatistics)。由于各研究领域的侧重点不同,我国统计界通常把生命科学实验研究中的统计学内容称为生物统计学,把医学研究中的统计学内容称为医学统计学,我国医学统计界又把预防医学研究中的统计学内容称为卫生统计学。随着医学研究模式的改变,医学领域各个学科相互渗透,所涉及的统计学研究工作已难以区分它们之间的差别。因此本书所述的统计学内容已基本涵盖了卫生统计学的内容。事实上,我国医学统计学工作者与国际统计界交流时也都统称为 Biostatistics。

第二节 统计学中的几个基本概念

一、变量和资料

在医学研究中,先根据研究目的确定研究对象,然后对研究对象的某项目或研究指标进行观察(或测量),这种观察项目或研究指标称为**变量**(variable),变量取值表示观察(或测量)结果或对应的观察结果,亦称资料(data)。例如调查某地 10 岁儿童性别与身高情况,则调查的对象是 10 岁儿童。调查项目为性别和身高。性别变量取值 1 表示被调查对象是男孩,取值 0 表示被调查的对象是女孩;身高变量取值为被调查对象的身高实际测量值/cm。在医学研究中,绝大多数观察(测量)指标在观察前是无法知道结果的,即观察结果是随机的。这种观察(测量)指标称为

随机变量(random variable),在医学统计书中经常简称为变量。随机变量可以分为连续型变量(continuous variable)和离散型变量(discrete variable)。

连续型变量的取值范围是一个区间,它可以在该区间中连续取值,即连续型变量可以取到区间中的任一值,并且一般有度量单位,观察连续型变量所得到的数据资料称为计量资料(measurement data),也可以称为定量资料。例如,用变量 X 表示 7 岁男孩体重,在测量精度理想的情况下,某一个区间中的任一数值都可能是 7 岁男孩的体重且度量单位为 kg,故体重变量 X 是一个连续型变量。观察 110 名 7 岁男孩的身高,可得到 110 个身高测量值/cm 的数据所构成的计量资料(详见例 2.1)。计量资料的分析可以先用第二章统计方法进行统计描述,然后根据资料特点选用合适的统计方法作进一步分析。

离散型变量取值范围是有限个值或者一个数列构成的,表示分类情况的离散型变量又称为分类变量。根据类别的有序性,分类变量又可以分为有序分类和无序分类。

无序分类(unordered categories)指变量取值仅表示互不相容的类别或属性。包括:① 二分类资料。如检查中学生是否近视眼,以每个学生为观察单位,检查结果可以是近视眼($X=1$ 表示),也可以不是近视眼($X=0$ 表示)。② 多分类资料。如果考察某人群的血型,以每个人为观察单位,检查的可能结果为 A、B、AB 和 O 型,也可以用 $X=0,1,2,3$ 分别表示。显然 X 的取值仅是起指示分类的作用,其数值大小并无实际意义。无序分类资料的分析应先按类统计汇总,统计每一类的观察单位数,并将按类汇总的统计结果编制成表格形式的资料,这种汇总后的资料又可称为计数资料,这类资料常用第六章的统计方法进行分析。

有序分类(ordinal categories)指变量取值不仅表示互不相容的类别而且表示各类在研究背景意义下的等级顺序,因此具有"半定量"意义。所以观察有序分类变量所得资料又称为等级资料。例如,患者治疗的可能结果有治愈、好转、有效、无效或者死亡。从治疗效果评价的角度上考察,这些分类是优劣等级的区别,因此这样的资料是有序分类的。研究者可以用 $X=0,1,2,3,4$ 分别表示以上等级,但等级之间的差别可以是量的差别,也可以是质的差别,这种差别有时难以精确度量。有序分类资料的分析也应先按类统计汇总,统计每一类的观察单位数,将分类统计结果编制成表格形式的资料,并把等级资料转化为秩,用第八章的统计方法进行分析。

有些观察指标,例如白细胞计数,其取值虽然是离散的,但不具有分类的性质,因此通常把这类观察指标的资料作为较为特殊的计量资料,并根据实际情况,选用第二章、第四章、第七章或第八章的统计方法进行统计分析。

在实际研究中,由于研究目的和结果解释的原因,有时需要把计量资料转换为分类资料。如血压测量值为计量资料,但若将其改记为高血压、正常血压和低血压,则血压测量值资料转换为血压有序分类资料。

二、个体(individual)

个体是统计分析根据研究目的所确定的最基本的研究对象单位,所以个体又称为**观察单位**。根据不同的研究目的,个体可以是一个人、一只大白鼠、一个家庭、一个地区、一个检测样品、一个采样点等。例如,观察单位是一个人,则 100 个观察单位就是 100 个人;又如观察单位是检测样品,则 50 个观察单位就是 50 个检测样品。

具有相同性质的观察单位称为**同质的**(homogeneous),否则,称为异质的(heterogeneous)。例如调查某地 1995 年正常成年女子的糖化血红蛋白(HbA1C),则研究对象是该地 1995 年的正

常成年女子，观察单位是每个女子，观察指标是糖化血红蛋白，观察值是每个人的糖化血红蛋白测量值，同一地区、同一年份、同为正常成年人和同为女性构成了研究对象同质的要素。同质的要求与研究目的有关，如调查某地1995年正常成年女子的雌性激素水平，尽管研究对象是同一地区、同一年份、同为正常成年人和同为女性，由于女性在绝经后的雌性激素水平有较大下降，如果研究者把绝经和未绝经的研究对象混合在一起，并且不加区分，则对于研究雌性激素水平而言，这些研究对象显然是异质的。

三、总体和样本

总体(population)是同质的所有个体某指标观察值(测量值)的集合。例如，研究目的是考察某时某地区10岁正常发育男孩的身高分布情况，则观察单位为每个男孩，观察指标(变量)为身高，观察值是每个男孩的身高测量值，同质个体的依据为同一地区、同一时间的10岁正常发育男孩。因此，在这时的该地区所有10岁正常发育男孩就是这个研究目的所确定的所有同质个体，而该身高总体就是该地区所有10岁正常发育男孩身高测量值的集合。由于在实际研究中，往往需要观察或测量多个指标，而这些指标之间往往伴有某种关联，故多个观察指标构成了个体的一组观察指标。为了叙述方便，往往简单地称总体是根据研究目的确定同质个体的全体。

总体分为有限总体(finite population)和无限总体(infinite population)。有限总体中个体总数是有限的；无限总体中个体总数是无限的。例如：研究某地区某时10岁正常发育男孩的身高分布情况。由于该地区在某时的10岁正常发育男孩的个数肯定是一个有限的数，因此该总体为有限总体。在医学研究中，许多研究目的所对应的总体在严格意义下是有限总体，但是这些总体中的个体总数往往非常大，与无限总体在实际应用中几乎无差别，所以把个体总数非常大的有限总体近似视为无限总体。

在医学研究中，总体往往非常大而很难或根本无法得到全部个体的观察值，因此需要通过抽样研究了解总体情况。抽样研究通常是在一个较大范围的研究对象中随机抽出一部分个体进行观察或测量，这些个体的测量值构成的集合称为**样本**(sample)，样本中的个体总数称为**样本量**(sample size)。通过样本资料的统计分析可以了解总体基本情况或推断总体的某些特征。在抽样研究中，随机抽出一部分个体进行观察或测量的过程称为**随机抽样**(random sampling，详见第十三章)，因此总体和样本的关系为：总体⊃样本。

刻画总体特征的指标称为总体参数(parameter)。例如总体中某个指标的所有个体观察值的平均数称为总体均数；如某研究的总体为全体正常人，则所有正常人中HbsAg呈阳性的比例称为HbsAg的总体阳性率或HbsAg的总体阳性发生率。刻画样本特征的统计指标称为**统计量**(statistic)，例如样本中某个指标的所有观察值的平均数称为样本均数；如某样本中HbsAg呈阳性的比例称为HbsAg的样本阳性率。在实际工作中，总体参数往往是未知的，这些未知总体参数可以用样本统计量进行估计。例如用样本均数估计总体均数、用HbsAg的样本阳性率估计HbsAg的总体阳性率等。

四、频率和概率

1. 随机事件

随机现象的某个可能观察结果称为一个随机事件，并通常用英文字母 A，B，C 等表示。如观察某一医院某一医生用某药抢救中风患者这一现象，被救活是一个可能的结果，而因抢救失败

而死亡是另一个可能的结果，抢救前不能确定何种结果将会出现，故被救活是一个随机事件（用 A 表示）。如果抢救的结果为患者被救活了，则称随机事件 A 发生了，抢救失败而死亡则称 A 未发生。

2. 频率（frequency）

用随机事件 A 发生表示观察到某个可能的结果，若在 n 次观察中，随机事件 A 发生了 m 次，则称 A 发生的比例 $f=\frac{m}{n}$ 为频率，m 称为频数。显然有 $0\leqslant f\leqslant 1$。在医学上所说的患病率、病死率等都是频率。如治疗了 n 个幽门螺旋杆菌感染（HP 阳性）的患者，其中有 m 个人治愈（HP 呈阴性），则治愈率 $f=\frac{m}{n}$。

频率 f 是一个统计量，由于个体的变异性，频率 f 呈一定的随机波动。如在某地区随机抽样调查糖尿病的患病率，其结果见下表

某地区糖尿病患病率的抽样调查结果

抽样（调查）人数 n	100	500	1 000	5 000	10 000	50 000	100 000	1 000 000
糖尿病人数 m	12	48	102	493	992	4 999	10 003	99 999
频率（患病率）f/%	12.00	9.60	10.20	9.86	9.92	10.00	10.00	10.00

由上述表可以看到频率 f 呈某种随机性。但随着抽样人数 n 的增大，频率（患病率）f 随机波动的幅度越来越小并且趋向常数 10%。可以证明：当观察次数 n 越来越大，频率 f 的随机波动幅度越来越小，并最终趋向于一个常数，这个常数被称为随机事件 A 发生的概率（又称为概率的统计定义）。

3. 概率（probability）

概率刻画随机事件发生可能性大小，其取值界于 0 和 1 之间。随机事件发生的可能性越小，概率越接近 0；随机事件发生的可能性越大，概率越接近 1。特别，不可能事件发生的概率等于 0，必然事件发生的概率等于 1。

在统计学中，如果随机事件发生的概率小于或等于 0.05，则认为是一个小概率事件，表示该事件在大多数情况下不会发生，并且一般认为小概率事件在一次随机抽样中不会发生，这就是小概率原理。小概率原理是统计推断的基础。

随机事件 A 发生的概率是一个总体参数，在许多情况下是未知的，实际工作中往往用频率。

五、个体变异和资料分布

同质个体的某指标之间的差异称为个体变异（individual variation）。在生物医学研究领域里，个体变异是普遍存在的，即使在完全相同的自然条件下生长的两个生物体也存在差异。如同为正常发育的 10 岁男孩，各人身高各不相同；又如，病情相同的患者服用相同的药物，其疗效也不尽相同。由于在观察或测量前都不能事先确定这些个体变异的大小，因此这些个体变异是随机的，而同一类的个体变异在概率意义下是有规律的，这种随机变异的规律性表现为观察值出现在不同范围中的概率，所以称这种随机变异的规律性为该观察指标（变量）取值的概率分布，简称为资料的分布。例如，观察某地区 10 岁健康男孩身高的分布情况，把身高分为 3 段：第一段为身高小于 125cm；第二段为身高在 125～135cm；第三段为身高高于 135cm。对于在该地区随机抽

一个10岁健康男孩并测量他的身高而言，该男孩身高在这3个范围中的任何一个都是可能的，所以在抽样前不能断定所抽到的健康男孩身高在哪个范围中。但如果抽查该地区许多10岁健康男孩的身高，统计这些男孩身高出现在这3个范围中的频率，并用这些频率作为相应概率的近似值，就可以得到身高分布的近似概率。例如，在该地区抽了10 000个10岁健康男孩并测量其身高，结果为身高小于125cm共有720人（占7.2%）；身高在125～135cm范围中共有8 950人（占89.5%）；身高大于135cm共有330人（占3.3%）。可以发现该地10岁健康男孩的身高规律：身高小于125cm的概率约为7.2%、身高在125～135cm范围中的概率约为89.5%以及身高大于135cm的概率约为3.3%。故对于随机考察的一个10岁健康男孩身高而言，虽因为随机性而不能断定其身高在哪个范围中，但可以肯定身高在125～135cm范围中的机会要远高于其他身高两个范围。本例只是一种较简单的概率分布，事实上，**任何随机现象或随机变异都有其固有的分布规律，即概率分布，在大量重复观察的条件下就会呈现其规律性。**详见第二章。

六、抽样误差和测量误差

在医学研究中，许多总体指标是未知的，需要用相应的样本统计量对其进行估计。由随机抽样造成的样本统计量与总体指标之间的差异称为抽样误差（sampling error）。抽样误差由个体变异和抽样所致，虽然在一次抽样研究中的抽样误差是随机的，但抽样误差在概率意义下是有规律的，这种规律称为抽样分布。在同一总体中进行大量重复抽样可呈现这种抽样误差的规律。由于个体变异普遍存在，所以抽样误差是不可避免的，但可通过增大样本含量减小抽样误差。

测量误差是指实际观察值呈现规律性地偏离观察真实值。如试剂原因使某些生化测量值普遍低于或高于实际真实情况，因仪器未校正而进行测量以致观察值普遍低于或高于实际真实值等，这些误差属于测量误差。测量误差可以通过改进措施消除或减少，所以这类误差是有可能控制的。如某些生理指标有上午的观察值高于下午的观察值的规律，若忽略观察时间的控制，则这些生理指标的观察值必将出现较大误差，但若在同一时间段观察这些生理指标，其观察结果的误差将大大减小，故这类误差也是可以控制的。

第三节　医学研究中的统计问题

一、医学统计学在医学研究过程中的作用

医学研究中存在大量的随机现象，为找到随机现象的规律性，需要大量重复地观察。但在医学研究的实际工作中往往难以实现。运用统计学方法可以从较少量的重复观察结果中找到随机现象的规律性。为达到此目的，在医学研究的各步骤中都需要用到统计学知识。如在研究设计时根据研究问题确定主要效应指标、主要研究因素、研究对象以及入选标准和排除标准，并考虑如何减少测量误差，如何估计样本量，如何减少或控制偏倚（bias），制定抽样方案或随机分组方案；进行数据管理和质量控制；对资料进行统计分析，并做出恰如其分的推断。所以在医学研究工作中，统计工作者应自始至终参与和解决与统计相关的问题。

二、医学研究中常见的统计学问题

由于个体变异的原因，许多观察结果具有不确定性，因此需要根据统计学原理进行科学设计

和统计学分析。

例如，为了研究正常人群的某个激素水平情况，需要根据研究问题制定研究对象的入选标准和排除标准，根据实际情况和研究问题的需要确定抽样范围和样本量。通过抽样得到样本资料后，用第二章所述的统计方法了解资料的分布情况，用第四章的统计方法对总体参数进行估计，这样就可以了解该激素在正常人群的总体分布概况和总体平均水平。

例如，某医生用两种降血脂药物治疗高血脂患者，用A药治疗20人，其中15人/75%有效，用B药治疗20人，其中17人/85%有效，能否根据这个结果推断B药优于A药？如果各治疗100个患者，B药的有效人数是否仍高于A药的有效人数？若治疗更多的人，结果是否会改变？改变的可能性有多大？对于这些问题的统计推断需用第六章的统计方法进行统计分析，才能得出相应结论。

例如，对于评价不同化疗方案对延长肿瘤患者生命的疗效，往往需要用第十章和第十一章所述的统计学方法对接受不同化疗方案患者的随访资料进行生存分析。

又如，在研究骨质疏松的患病率的社区调查研究中，常常需要通过测量骨密度进行诊断是否患骨质疏松病。但测量骨密度费用比较贵而且测量也不方便。由于骨质疏松症的另一特点为在相同身高、肥胖程度、性别和年龄的情况下，骨质疏松患者的体重低于正常人的体重，因此可以利用这一重要特征，将明显不可能患骨质疏松症而无需做骨密度检查的调查对象筛选出来。因此需用第九章和第十一章的统计学方法，对已收集正常人的体重、身高、臀围、腰围、性别和年龄等资料进行统计分析，得到体重近似估计表达式：

$$\text{体重} = a + b_1 \times \text{身高} + b_2 \times \text{腰围} + b_3 \times \text{臀围} + b_4 \times \text{性别} + b_5 \times \text{年龄}$$

然后用该体重近似估计表达式对人群进行筛选：用调查对象的身高、腰围、臀围、性别和年龄代入体重近似表达式中，计算出相应正常人的体重近似估计值。如果调查对象的实际体重远高于调查对象的体重估计值，则可以基本认定该对象不可能患骨质疏松症，因此没有必要进行骨密度检查。这样一部分调查对象可以不必作骨密度检测，从而节省了相关的费用。

第四节 学习医学统计学应注意的问题

医学统计学是医学生的必修课。根据不同专业的特点，相应的课时数可能有所增减。但无论是预防医学专业的学生、临床医学专业的学生或其他医学专业的学生，都将会在今后从事医学研究和实际工作中遇到各种各样的医学统计问题，因此掌握必要的医学统计学知识是适应今后工作的一种基本要求。为此，在学习本课程时，应注意：

(1) 正确理解医学统计学中的一些基本概念，并注意体会这些基本概念在不同实际研究问题中的含义。

(2) 要清楚地认识到：在大量重复观察的意义下，随机现象的观察结果所呈现的某种规律反映了总体的特征，并体会这种规律在实际研究问题中的含义。

(3) 要较好地理解：统计分析的目的是通过样本的信息了解总体的特征。并体会“通过样本的信息了解总体的特征”在实际统计分析中的含义。

(4) 资料分布(变量的概率分布)、抽样误差和假设检验是统计学最重要的内容，也是能否学习好医学统计学的关键处。本书在有关练习中给出了一些随机模拟习题，这是帮助学生理解抽

样误差、概率分布和假设检验基本原理的最好方法之一，希望每位学生在有条件的情况下尽量完成这些习题。

(5) 掌握研究设计的基本技能。本课程不仅要求学生掌握研究设计的基本原则，更重要的是要求学生能自己从事一些简单的研究设计。这样可以增强学生的动手能力、帮助学生较好理解研究设计的原理和体会医学统计学在实际研究工作中各个环节的作用和影响。为此本书给出了3个不同医学研究领域的完整实例，使学生可以较好地感悟实际医学研究工作中的医学研究设计、随机抽样、随机分组和统计分析策略等环节中的具体情况和处理措施。

(6) 熟练使用统计软件也是学习好医学统计学的一个基本要素。在实际工作中，绝大多数的统计分析工作是使用有关统计软件实现的。本书较详细地介绍了Stata软件的基本使用方法，同时在光盘中给出了SPSS软件和SAS软件的相关使用方法。建议学生至少能熟练掌握一个统计软件的使用，这样可以增强实际动手能力。但需要注意的是：统计分析≠统计计算。统计软件只是提供了统计计算的一个工具，从而减轻统计计算的压力，但统计软件不能代替统计工作者的思维。学生不仅应注重掌握数据统计分析方法的选择，而且应注重分析结果的解释，从而提高学生分析问题和解决问题的能力。

综上所述，正确理解基本概念、掌握研究设计和统计分析方法、提高实际动手能力并在实践中进一步加深理解是学好医学统计学的关键。系统地了解在实际研究过程中与统计学相关的问题是用好统计学的前提。只有把学习医学统计学与实际医学研究工作结合起来，才能真正理解医学统计学基本概念和掌握医学统计学分析方法。

小　结

总体是同质的所有个体某指标观察值(测量值)的集合。为了叙述方便，往往简单地称总体是根据研究目的确定同质个体的全体。

某个观察指标的总体均数就是总体中所有个体的该观察指标测量值的平均数。

随着观察次数 n 增大，随机事件 A 发生的频率 f 趋向于 A 发生的概率。

绝大多数的医学研究观察结果具有个体变异和随机性，在大量重复观察的意义下都会呈现一定的规律性，即资料分布(亦称资料的概率分布或资料的总体分布)。统计分析的目的就是通过收集和分析样本资料了解总体。

变量分为连续型变量和离散型变量。连续型变量取值所对应的资料称为计量资料；具有分类意义的离散型变量取值所对应的资料分为无序分类资料(整理汇总成表格形式的资料亦称计数资料)和有序分类资料(整理汇总成表格形式的资料亦称为等级资料)；另有一部分资料虽属于离散型变量的取值，但其不具有分类性质，故把这类资料作为较特殊的计量资料。

习　题

1. 某医生收治200名患者，随机分成2组，每组100人。一组用A药，另一组用B药。经过2个月的治疗，A药组治愈了90人，B组治愈了85名患者，请根据现有结果评议下列说法是否正确，为什么？

(1) A药组的疗效高于B药组(即两个样本的疗效)。

(2) A 药的疗效高于 B 药(即两个人群的疗效)。

2. 某校同一年级的 A 班和 B 班用同一试卷进行一次数学测验。经过盲态改卷后,公布成绩:A 班的平均成绩为 80 分,B 班的平均成绩为 82 分,请评议下列说法是否正确,为什么?

(1) 可以称 A 班的这次考试的平均成绩低于 B 班,不存在抽样误差。

(2) 不能推断 A 班的这次考试的平均成绩低于 B 班,存在抽样误差。

(3) 通过这次考试的平均成绩,说明 B 班的数学平均水平高于 A 班。

(4) 对于评价两个班级的数学平均水平而言,这次考试成绩只是一次抽样观察结果,所以存在抽样误差,不能仅凭这次考试的平均分差异推断两个班级的平均水平的高低。

(5) 对于研究两个班级的这次考试成绩而言,A 班所有学生的这次考试成绩构成了一个总体 A,B 班所有学生的这次考试成绩构成了一个总体 B。

(赵耐青)

第二章　统计描述

本章将叙述资料的统计描述。搜集、整理资料后首先要考虑表达资料，使人们对搜集到的资料有一个大致的了解，为此需对资料作统计描述。资料的统计描述一般可用统计表、统计图、统计指标。不同类型的数据资料有不同的描述方法。本章第一节将讲述计量资料的统计描述，第二节讲到的概率分布是常用来描述资料分布规律的三种随机变量分布，第三节讲述计数资料的统计描述，第四节讲述常用的统计图。

第一节　计量资料的统计描述

一、统计图表

描述计量资料分布规律的统计图表主要是频数表与频数(频率)图。

(一) 频数表

1. 频数表的编制

连续变量与离散变量的频数表的制作方法略有不同，下面各举 1 例说明之。

例 2.1　某市 1995 年 110 名 7 岁男童的身高/cm 资料如下，请作统计描述。

121.4	119.2	124.7	125.0	115.0	112.8	120.2	110.2	120.9	120.1
125.5	120.3	122.3	118.2	116.7	121.7	116.8	121.6	120.2	122.0
121.7	118.8	121.8	124.5	121.7	122.7	116.3	124.0	119.0	124.5
121.8	124.9	130.0	123.5	128.1	119.7	126.1	131.3	123.8	116.7
122.2	122.8	128.6	122.0	132.5	122.0	123.5	116.3	126.1	119.2
126.4	118.4	121.0	119.1	116.9	131.1	120.4	115.2	118.0	122.4
120.3	116.9	126.4	114.2	127.2	118.3	127.8	123.0	117.4	123.2
119.9	122.1	120.4	124.8	122.1	114.4	120.5	120.0	122.8	116.8
125.8	120.1	124.8	122.7	119.4	128.2	124.1	127.2	120.0	122.7
118.3	127.1	122.5	116.3	125.1	124.4	112.3	121.3	127.0	113.5
118.8	127.6	125.2	121.5	122.5	129.1	122.6	134.5	118.3	132.8

解：本例的观察指标身高属连续变量，其频数表的制作需先划分组段。为保证收集到的所有数据均在所划的组段内，第一组段的下限必须小于等于全部数据中的最小值，最后一个组段的上限必须大于等于全部数据中的最大值。

组段划分的具体步骤如下：

(1) 确定组数　制作频数表是为了较好地显示出数据的分布规律，故组数不宜过多，但也不能太少。因为组数太少也会掩盖数据分布的规律。适宜的分组数与数据个数 n 的多少有关，大

体上当数据个数 n 在 30 左右时，可分 5 到 6 组，随着 n 的增加，分组数适当增加。较大样本时，一般取 10 组左右。

(2) 确定组距 组距可以相等，也可以不相等。实际应用时一般采用等距分组，其组距≈极差/组数。极差是全部数据中的最大值与最小值之差，本例最大值为 134.5cm，最小值为 110.2cm，故极差 $R = 134.5\text{cm} - 110.2\text{cm} = 24.3\text{cm}$。因为 $24.3/10 = 2.43 \approx 2$，故取组距 = 2。

(3) 确定各组段的上下限 每个组段的起点被称为该组的下限(lower limit)，终点被称为上限(upper limit)，上限 = 下限 + 组距。显然第一组段必须包含最小值，其下限一般取包含最小值的、较为整齐的数值。本例最小值为 110.2，故取 110 为第一组段的下限，其上限 = 110 + 2 = 112。值得注意的是各组段不能重叠，故每一组段均为半开半闭区间。表 2.1 第(1)栏中的"110～"表示区间[110,112)，"112～"表示区间[112,114)，依此类推。

划分组段后，统计出各组段内的数据个数(频数)，即得频数表。本例的频数表见表 2.1 第(1)、(2)栏。

表 2.1 110 名 7 岁男童身高频数分布

身高组段 (1)	频数 (2)	频率/% (3)	累计频数 (4)	累计频率/% (5)
110～	1	0.91	1	0.91
112～	3	2.73	4	3.64
114～	4	3.64	8	7.27
116～	10	9.09	18	16.36
118～	15	13.64	33	30.00
120～	22	20.00	55	50.00
122～	21	19.09	76	69.09
124～	14	12.73	90	81.82
126～	10	9.09	100	90.91
128～	4	3.64	104	94.55
130～	3	2.73	107	97.27
132～	2	1.82	109	99.09
134～136	1	0.91	110	100.00
合 计	110	100.00	—	—

例 2.1 中的指标身高是一个连续变化的量。这种连续变化的变量被称为连续型变量。已婚育龄妇女的现有子女数，幼儿的牙齿数等却不然，其取值是 0、1、2 等不连续的量，这种变量被称为离散型变量。上面所讲的频数表的制作方法是针对连续型变量的数据资料而言的。对于离散型变量其频数表的编制较为简单，其每一组段往往是一个取值。如表 2.2 的第(1)、(2)栏。

2. 频率与累计频率 频数表中的各组频数之和等于总例数 n，将各组的频数除以 n 所得的比值被称为频率。频率描述了各组频数在全体中所占的比重，各组频率之和为 100%。表 2.1 与表 2.2 的第(3)栏为相应的频率。

由于各组段频数与总例数有关，当几个样本资料的总例数不等时，难以根据各组段频数多少比较他们的数据分布的规律，但可利用频率进行比较。

在实际应用中往往需知道在某个指定值以下的数据频数或频率，这种频数或频率被称为累