公共卫生硕士(MPH)系列教材

总主编 姜庆五

医学统计学基础

（第二版）

余金明 凌 莉 主编

YIXUE TONGJIXUE JICHU

复旦大学出版社
www.fudanpress.com.cn

公共卫生硕士(MPH)系列教材

总主编 姜庆五

医学统计学基础

（第二版）

● 余金明 凌 莉 主编

副主编 贺 佳 高 歌 刘玉秀

主 审 曹素华

编写者（以姓氏笔画排序）

邓 伟（复旦大学）

叶小飞（第二军医大学）

刘玉秀（南京军区总医院）

刘 浩（复旦大学）

余金明（复旦大学）

罗剑锋（复旦大学）

贺 佳（第二军医大学）

秦国友（复旦大学）

凌 莉（中山大学）

高 歌（苏州大学）

董 英（上海中医药大学）

学术秘书 董 英（兼）

复旦大学出版社

图书在版编目(CIP)数据

医学统计学基础/余金明,凌莉主编. —2 版. —上海:
复旦大学出版社,2009.11
(公共卫生硕士(MPH)系列教材)
ISBN 978-7-309-06960-0

Ⅰ. 医… Ⅱ. ①余…②凌… Ⅲ. 医学统计-研究生-教材 Ⅳ. R195.1

中国版本图书馆 CIP 数据核字(2009)第 200075 号

医学统计学基础(第二版)
余金明 凌 莉 主编

出版发行 復旦大學出版社 上海市国权路 579 号 邮编 200433
86-21-65642857(门市零售)
86-21-65100562(团体订购) 86-21-65109143(外埠邮购)
fupnet@fudanpress.com http://www.fudanpress.com

责任编辑 宫建平
出 品 人 贺圣遂

印 刷 上海市崇明县裕安印刷厂
开 本 787×960 1/16
印 张 22.25
字 数 412 千
版 次 2009 年 11 月第二版第一次印刷

书 号 ISBN 978-7-309-06960-0/R·1120
定 价 48.00 元

内 容 提 要

《医学统计学基础》是一本用于公共卫生硕士（MPH）专业学位教学的教材，内容包括医学统计学的基本概念、基本原理和方法、计量资料和计数资料的统计描述、单变量的统计推断、研究设计概述等。

为帮助读者更便捷地掌握统计分析方法的正确选择和有效提高数据分析的技能，本教材的章节编排以研究设计的类型和资料性质为依据，各章以实际研究资料为案例，引出针对该类研究资料的相关统计分析方法，介绍相关统计分析方法的基本概念、知识要点、分析过程、应用条件和常见误用等。同时介绍常用的统计分析软件，如SPSS、Excel、STATA和SAS等的基本操作，以及计算机分析结果的正确解释，淡化统计公式和统计计算过程。本教材非常方便读者针对自己实际分析资料来查找正确的统计分析方法和进行合理的结果解释。为方便读者复习巩固理论知识，每章都附有复习思考题及参考答案。

余金明　复旦大学公共卫生学院流行病与卫生统计学教授，博士生导师。1986年毕业于安徽医科大学卫生系，获学士学位；1991年获流行病学硕士学位；1996年毕业于上海医科大学，获流行病学博士学位。1995～1996年和1998年分别在荷兰鹿特丹ERASMUS大学、比利时Leopold王子热带医学研究所做高级访问学者。现任复旦大学公共卫生学院卫生统计教研室副主任、临床流行病学研究中心副主任，国家级卫生应急专家，卫生部、教育部、科技部、上海市专家库专家，教育部高等学校预防医学专业教学指导分委会委员，上海市卫生统计学专业委员会委员，上海市流行病学专业委员会委员，中国医师协会循证医学专业委员会委员，中国医师协会心血管内科医师分会委员，中国医师协会循证医学专业委员会临床试验学组委员，中国动脉硬化学会诺美基金专家组成员，复旦大学生物安全教育部重点实验室专家组成员，*BioMed Central*、*Lancet*、《预防医学杂志》、《中华流行病学杂志》、《中华内科学杂志》、《中华内分泌代谢杂志》、《中华心血管病杂志》、《中华高血压病杂志》、《中国动脉硬化杂志》、《复旦大学学报》等特约审稿专家，《疾病控制杂志》、《环境与职业医学杂志》、《中国循证心血管医学杂志》等编委。

凌　莉　中山大学公共卫生学院医学统计学与流行病学系教授，博士生导师。1985年和1988年于华西医科大学分别获预防医学学士学位和卫生统计学硕士学位；2001年毕业于中山大学，获卫生统计学博士学位。1996年、2007年分别在日本National Institute of Public Health和美国Yale University做访问学者。现任中国卫生信息学会理事、中国卫生信息学会医学统计教育专业委员会秘书长，广东省卫生统计学会副秘书长，广东省现场统计学会常务理事，国家食品药品监督管理局（SFDA）新药统计学评审专家。曾任全国高等学校临床医学专业八年制卫生部规划教材《医学统计学》（第二版），全国高等医药院校研究生规划教材《医学统计学》（第三版），普通高等教育“十一五”国家级规划教材《医学统计学》（第二版），普通高等教育“十一五”国家级规划教材《卫生统计学》，教育部研究生工作办公室推荐研究生教学用书《医学统计学与电脑实验》（第二版和第三版），研究生教学用书《生物医学研究的统计方法》等的编委。

序　言

公共卫生硕士(MPH)是由国务院学位委员会批准设置的一个新的专业学位。MPH将成为公共卫生人才培养的重要职业教育形式。

MPH学位教育的目的是培养高层次卫生管理与疾病预防应用型人才。要求MPH的学生具备广博的专业知识、创新性的科学思维;勇于开拓、善于实践;能胜任卫生行政部门、医疗机构、疾病控制及卫生监督部门的高层次卫生管理与疾病预防的重要工作。在MPH学位教育过程中,应该注重拓宽学生的知识面,注重现代科学技术知识的掌握,重点培养学生分析问题和解决问题的能力。

复旦大学公共卫生学院在MPH学位教育过程中,注重理论与实践相结合,课堂教学与课题研究相结合,教材建设与教学实践相结合。2002年我们组织编写和出版了第一版公共卫生硕士系列教材,收到良好的教学效果,大大提升了MPH教学质量,为MPH学位培养作出了贡献。

根据近7年教材的使用情况及MPH学位教学改革的需要,我们再次策划和组织了MPH系列教材的第二版,包括MPH学位的必修课,以及根据学生各自的基础和知识结构确立的选修课。此系列教材包括:①卫生事业(保健)管理专业方向的课程,如卫生政策分析、卫生服务研究、卫生经济学、医疗保险学、医院管理、社区卫生管理与评价等。②流行病学与疾病控制专业方向的课程,如流行病学基础、流行病学方法、现场流行病学、医学统计学基础、数据分析方法、现场调查技术等。③环境医学与卫生监督专业方向的课程,如环境医学导论、灾害医学、营养与健康、食品安全、卫生监督等。④妇儿保健与健康促进专业方向的课程,如健康促进理论与实践、卫生服务理念与方法等。

MPH学位培养正在我国蓬勃发展,此套教材是我们开展MPH学位教育的探索与尝试,若有不当之处,敬请读者批评指正。我们将与全国的公共卫生教育者一起,为开拓与完善我国MPH学位教育与教材建设作出贡献。

姜庆五

2009年11月

第二版前言

公共卫生硕士(MPH)专业学位教育是应用型的硕士研究生教育，其对象主要为医学卫生领域从事本专业的工作人员。虽然他们都毕业于公共卫生与预防医学或临床医学等相关专业，有一定的医学统计学基础，但由于工作多年，对医学统计学的基础内容有所生疏。故本书将沿第一版的基本思路，仍从一些基本概念讲起，但教材内容将突破传统卫生统计学教科书编排模式，强调统计学方法在实际统计分析中的正确应用。前5章为统计学的基础知识和基本概念，自第六章起将以研究设计为主线，针对此类研究设计的资料特征，正确选择统计分析方法进行统计分析。由于统计分析方法的正确运用与研究设计密不可分，故在第一版的基础上增加了“研究设计概述”一章。本教材为医学统计学基础，内容主要涉及单变量统计学方法，多变量统计学方法将由另一部教材详细介绍。

本教材注重培养学生分析问题和解决问题的能力，注重理论与实践相结合。各章节以案例为引入点，分析问题所在，导入相关基本概念、理论和方法；然后回过头再看案例分析方法的正确选择，将涉及的主要统计概念和核心内容以知识要点的形式给出，便于掌握精髓和复习；最后给出新的案例，供学生应用以上所学知识。

本教材淡化了统计公式的推导和统计计算过程，重点放在基本概念和基本原理，以及统计方法的用途、适用条件、应用注意事项及结果的解释与表达等方面。

在编写内容和结构上，考虑到教师授课的实际需要，使教师使用本教材组织教学时，既可按传统模式讲授，以案例作为补充，也可以案例为先导进行教学，使课堂讲解内容更加形象、生动。

多年来在与MPH学员的交流中获悉，针对基础统计，日常工作时更倾向于选择操作简便的统计分析软件，如Excel和SPSS，本书相关章节后都附上了所举案例的几种常用统计软件操作步骤及主要运算结果，并进行了简单解释。同时提供Excel、STATA、SPSS和SAS等统计分析软件的应用介绍，但以SPSS(11.0及以上版本)和Excel(Office 2003及以上版本)为主。

本教材编写过程中，主审曹素华教授倾注了大量的心血，各位作者在百忙中抽出宝贵时间参与本书的编写；还得到复旦大学公共卫生学院院长姜庆五教授、书记傅华教授、公共卫生硕士(MPH)办公室单维良老师以及卫生统计教研室老师们的大力支持和指导，在此一并致谢。

由于编写者的水平有限及编写时间匆促，本书可能存在不少缺点和错误，恳请读者提出宝贵意见，以便进一步修改和完善。

余金明　凌　莉

2009 年 11 月

目　录

第一章 绪 论

第一节 医学统计学在医学研究中的应用

医学研究主要是针对人类的疾病与健康状况及其相关的各种影响因素。由于影响人体的因素错综复杂，许多观察结果都不能事先确定，即使条件完全相同的两次观察其结果往往也会不同。观察结果的这种不确定性，我们称为随机性。为了处理这种不确定性，透过偶然性来解释所观察到的现象并发现其中的规律性需要运用统计学知识。

统计学（statistics）是研究数据收集、整理、分析、推断等原理和方法的学科。它在医学研究中的运用已越来越广泛和深入，并逐渐形成了一门分支学科，即医学统计学（medical statistics）。医学统计学是统计学原理和方法在医学研究领域的具体应用。目前，许多国际性医学研究项目均需医学统计学人员参加。例如，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。医学统计学已经成为医学各专业本科生和研究生的必修课程。

医学统计学通过指导医学研究设计、资料收集和结果分析，以尽可能少的重复观察获取足够的信息量，有效利用有限的资料，作出精确可靠的结论。在生物制药领域，临床试验涉及的随机化分组、双盲设计、减少和控制偏移、估算样本量等一系列过程均需要运用医学统计学知识。在公共卫生领域，众多危险因素的分析、生存时间分析、疾病自然史模型等的应用都对医学统计学提出了越来越高的要求。

第二节　医学统计学中的几个基本概念

一、随机现象与随机事件

某类现象，在个别的实验或观测中呈现出不确定性，但在大量重复试验或观察中又具有统计规律性，我们称它为**随机现象**。随机现象的结果构成**随机事件**。如一个人可能患某种疾病，也可能不患某种疾病，具有随机性，而这个人最后患该病的结果则构成一个随机事件，若这个人未患该病则构成另一个随机事件。

二、随机变量

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，称为变量值(value of variable)或观察值(observed value)，亦称为**资料**(data)。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)，在医学统计学书中经常简称为变量。根据变量的类型，可分为连续型变量(continuous variable)和离散型变量(discrete variable)。

(一) 连续型变量

连续型变量又称**数值变量**(numerical variable)，对应的资料称为计量资料(measurement data)或定量资料(quantitative data)。连续型变量的取值是一个区间，即可以连续性取值，并且有一定的度量单位。如身高(cm)、体重(kg)和血压(mmHg)等。

(二) 离散型变量

离散型变量依其取值情况，可以分为具有分类性质的资料和不具有分类性质的资料。

1. *不具有分类性质的离散资料*　有些观察指标，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，通常把这类指标的资料按特殊的计量资料处理。

2. *具有分类性质的离散资料*　表示分类情况的离散型变量，其变量取值范围是有限个值，常称为分类资料(categorical variable)。分类资料根据是否有序又可分为有序分类资料和无序分类资料。

(1) **无序分类变量**(unordered categorical variable):是指所分类别或属性之间没有程度和顺序的差别。它又可分为:①二项分类,如性别(男、女)、药物反应(阴性和阳性)等;②多项分类,如血型(O、A、B、AB)、职业(工、农、兵、学、商)等。对于无序分类变量的分析,应先按类别分组,清点各组的观察单位数,编制分类变量的频数表,所得资料为无序分类资料,亦称计数资料。

(2) **有序分类变量**(ordinal categorical variable):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,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 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。

对于有序分类变量,往往先按等级顺序分组,清点各组的观察单位个数,编制有序变量(各等级)的频数表。

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定量资料,并根据实际情况,选用合适的统计方法进行统计分析。

在实际应用中,由于研究目的和结果解释的原因,人们有时需要将一种类型的变量转化为另一种类型。但变量只能由高级向低级转化,即定量>有序分类>二分类。不能作相反方向的转化。例如,血压测量值为定量资料,但若将其改记为高血压、正常高值和正常血压,则血压测量值的定量资料转换为血压有序分类资料。但需要注意,这种转换可能损失部分信息。离散型变量常常通过适当的变换或连续性校正后借用连续型变量或有序分类变量的方法来分析。

三、个体

个体(individual)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,个体又称为观察单位(observed unit)。根据不同的研究目的,个体可以是一个人、一只大鼠、一个家庭、一个地区,甚至一个检测样品、一个采样点等。根据研究目的确定的具有相同性质的观察单位称为**同质**(homogeneous),否则称为**异质**(heterogeneous)。根据研究目的所确定的所有同质个体某指标实际值的集合,称为总体。在实际工作中,一个个体往往需要观察一组指标,为了叙述方便,常简单地将总体作为根据研究目的所确定的同质所有观察对象的集合。如果观察单位异质,则不能归于一个总体。个体间存在差异是绝对的,这种现象称为变异(variation)。例如,同患某病的患者具有同质性,但他们的身高、体重又存在变

异。变异构成了统计研究的基础,没有变异就没有统计学。

四、频率与概率

一个随机试验在相同条件下重复进行试验时,个别结果看来是偶然发生的,但当重复试验次数相当大时,总有某种规律性出现。用随机事件 A 发生表示观察到某个可能的结果,若观察 n 次,随机事件 A 发生了 m 次,则称 A 发生的比例 $f = m/n$ 为**频率**(frequency),m 称为频数。在医学上所说的患病率、病死率等都是频率。

概率用来表示随机事件发生可能性的大小,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。不可能事件发生的概率为 0;必然事件发生的概率为 1。它们都具有确定性,可看作随机事件的特例。随机事件 A 发生的概率在许多情况下是未知的。在实际工作中,当概率不易求得时,只要观察单位数足够多,可以将频率作为概率的估计值。但在观察单位数较少时,频率的波动性较大,用于估计概率是不可靠的。

在统计学中,如果随机事件发生的概率$\leqslant$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。

五、总体与样本

根据研究目的而确定的同质观察单位的全体称为**总体**(population),它是同质的所有个体某项指标观察值(测量值)的集合。例如"治疗某病的药物疗效研究"问题中,总体是指所有使用该药某病患者的某疗效指标值的集合。又如,"全国 10 岁男童身高水平研究"的例子中,总体是指全国所有 10 岁男童的身高值的集合。

总体是研究对象组成的目标群体,有时候是假想的一个范畴。如上述"治疗某病的药物疗效研究"问题中,使用此药的某病患者,没有时间、空间的限制,观察单位可以无限多个,这种总体称为无限总体。

由于研究对象可能是无限总体,或者总体中观察单位数很大,或者观察方法对人体是有损害的等限制因素,导致我们无法对每一个个体进行观察,这时便需要从总体中抽取一部分的观察单位进行观察研究,然后根据抽取的这部分观察单位的信息来推断总体的特征。被抽取的这一部分观察单位就称为样本。从这个意义上来说,总体是我们研究所关心的目标群体,而样本是了解总体的手段和方法。样本所包含的观察单位数称为样本量。在上述"全国 10 岁男童身高水平

研究”的例子中，全国 10 岁男童的身高水平构成的总体是研究所关心的目标群体，通过一定的方法抽取的部分 10 岁男童的身高则构成了一个样本。

六、参数与统计量

参数是描述总体特征的统计指标。比如，全国 10 岁男童的身高平均数，即是用来描述全国 10 岁男童身高的平均水平。

统计量是描述样本特征的统计指标，是样本的函数。比如从全国各地抽取部分 10 岁男童，则样本中所有男童身高的平均数，即样本均数。不同样本有不同的样本均数，它随样本变化。在实际研究中，由于总体参数往往是未知的，所以需要通过相应的样本统计量来进行估计。

七、抽样误差及抽样分布

抽样研究中，研究的目的是了解总体的参数，而实际观察得到的只是统计量的观察值。为了使样本对总体具有代表性，样本代表的信息可以用统计方法来推断总体，样本的抽样需要符合随机化原则，即要求总体中的每个个体有均等的机会被抽取。

随机化抽取，可以使样本具有代表性，但也同时使抽取的样本具有随机性。所以，随机抽样获得的样本统计量与总体参数之间可能存在差异，以及不同次抽样的统计量观察值之间也可能存在差异。这种差异称为抽样误差。但是在多次独立重复抽样的情况下，抽样误差和样本统计量呈现出概率分布的规律。样本统计量的这种概率分布称为抽样分布。

（余金明）