



北京大学医学教材

*Health Statistics*

主编 王燕 康晓平

# 卫生统计学教程

北京大学医学出版社

國立臺灣大學圖書室

Principles of Statistical Inference

第二版

貝葉統計學教程

北京大學出版社

北京大学医学教材

# 卫生统计学教程

主 编 王 燕 康晓平

编写人员（按姓氏笔画为序）

王 燕 王洪源 任正洪 安 琳  
李 凯 易伟宁 罗树生 郑迎东  
高燕秋 康晓平

北京大学医学出版社

# WEISHENG TONGJIXUE JIAOCHENG

## 图书在版编目 (CIP) 数据

卫生统计学教程/王燕, 康晓平主编. —北京: 北京大学医学出版社, 2006. 3  
ISBN 7-81071-623-9

I. 卫... II. ①王...②康... III. 卫生统计—医学院校—教材 IV. R195.1

中国版本图书馆 CIP 数据核字 (2006) 第 010105 号

## 卫生统计学教程

---

主 编: 王 燕 康晓平

出版发行: 北京大学医学出版社 (电话: 010-82802230)

地 址: (100083) 北京市海淀区学院路 38 号 北京大学医学部院内

网 址: <http://www.pumpress.com.cn>

E - mail: [booksale@bjmu.edu.cn](mailto:booksale@bjmu.edu.cn)

印 刷: 莱芜市圣龙印务有限责任公司

经 销: 新华书店

责任编辑: 赵 曼 责任校对: 杜 悦 责任印制: 郭桂兰

开 本: 787mm×1092mm 1/16 印张: 18 字数: 452 千字

版 次: 2006 年 7 月第 1 版 2006 年 7 月第 1 次印刷 印数: 1—3000 册

书 号: ISBN 7-81071-623-9/R·623

定 价: 29.00 元

版权所有, 违者必究

(凡属质量问题请与本社发行部联系退换)

# 北京大学医学教材预防医学系列

## 教材编审委员会

主 任 胡永华

副主任 郝卫东

委 员 (按姓氏笔画为序)

王 生 王 燕 吴 明 宋文质

李 勇 李曼春 周宗灿 季成叶

胡永华 郝卫东 郭 岩 郭新彪

黄悦勤

秘 书 康凤娥

# 序

随着生命科学技术的日新月异,在我国高等教育体制改革的带动下,医学教育教学改革不断深入,医学教育逐渐由职业化教育转向具有职业特点的综合素质教育。着眼于 21 世纪,医学教育将更注重人才的综合培养,不仅要培养学生具有学科专业知识和能力,而且要具有知识面宽、能力强、素质高的特点,注重创新精神、创新意识、创新能力的培养。

1995 年以来,通过教育部、卫生部及北京市等各级教育教学改革项目的研究与实践,我校着力于人才培养模式和课程体系的研究,实现融知识、能力、素质于一体的综合培养,拓宽专业口径,特别强调理论与实践的结合,培养学生自学和创新的精神和能力,树立终身学习的观念;进行了课程内容、教学方法和考核方法的研究和实践;改革教与学的方法,以学生为主体,以教师为主导,引导学生主动学习,注意因材施教,注重加强人文素质的培养,强调在教学过程中的教书育人。

在改革实践中我们深刻认识到教材建设在教学过程中起着重要的作用。但长期以来医学教育一套教材一统天下的局面,未能充分体现各医学院校的办学特点,未能及时反映教学改革及教学内容的更新。为此我们邀请了北医及部分兄弟院校各学科的专家教授编写了这套长学制教材。

这套教材的编写工作力求符合人才培养目标和教学大纲,体现长学制教学的水平,探索和尝试突破原有教材的编写框架;体现北医教育观念的转变、教学内容和教学方法改革的成果和总体水平,确立以学生为主体的人才培养模式,有利于指导学生学习和思考,有利于训练学生临床思维的能力,培养学生的创新意识;体现教学过程中的“双语”教材要求,将学生必须掌握的词汇编入教材之中,部分教材配有英语专业词汇只读光盘。

本套教材汇集了北医及部分兄弟院校的专家教授们多年来积累的教学经验和教学经验,在编写中也进行了大胆的尝试。衷心希望该套教材的出版能为我国的医学教育贡献一份力量,使医学教育的教材建设能够百花齐放。但是由于学科专业发展的不平衡,教材中难免存在不足之处,欢迎有关专家学者批评指正。

韩启德

2002 年 7 月

# 前 言

《卫生统计学教程》是为预防医学七年制学生编写的教材。根据长学制教学要求，我们将《卫生统计学教程》内容分成两个部分：一部分是基本卫生统计学，另一部分是高级卫生统计学，主要是医学多变量统计。本教材主要介绍卫生统计学的基本内容，侧重于常见统计方法在预防医学中的应用，可以作为预防医学长学制学生本科学习阶段或预防医学五年制学生专业课学习阶段的教材。本教材的编写有以下两个方面的特点：

1. 强调基本理论和基本技能的实际应用。全书共 13 章，其中有 11 章介绍卫生统计学的基本理论、基本知识和基本技能，有 2 章介绍居民健康统计方法及其应用。为使学生在学习和理解统计方法原理的同时学会正确应用，各章编写重点突出了两个方面的内容：(1) 由医学研究和实践的实际例子引出各章的统计基本概念和方法，目的是让学生通过该章内容的学习后能够解决类似的实际问题；(2) 各章介绍的统计方法都给出相应的 SPSS 统计软件的计算程序和计算结果，目的是让学生理解统计概念的同时学会用 SPSS 统计软件快速计算出结果，用更多的时间去理解统计原理和解释结果。

2. 本书某些内容的安排与以往的《卫生统计学》教材略有不同。(1) 以往《卫生统计学》教材中都有独立的“非参数统计方法”一章，在教学实践中我们认为“非参数统计方法”和“参数统计方法”虽适用条件不同，但其欲解决的问题有相同之处，如果根据欲解决的问题类型将相应的“参数统计方法”和“非参数统计方法”放在一章可以起到对比和鉴别的作用，有利于学生今后的实际应用。因此本教材将原来的“非参数统计”一章拆开放到了相应的几个章节中。(2) 以往《卫生统计学》教材的“方差分析”一章都安排在“实验设计”这章前面，学生在理解“方差分析”时感到费解。为便于学生学习和理解，本教材将“实验设计”一章提到“方差分析”之前，并按“实验设计”中讲解的设计类型在“多组数值变量比较的假设检验”一章中介绍了相应的方差分析方法。

本书的编写人员都是北京大学公共卫生学院从事卫生统计学教学多年的中青年教师，积累了一定的教学经验，能体会学生初次学习《卫生统计学》的难点。本教程以表达正确、思路清晰、重点突出、易于理解、接近实际为原则进行编写，参考了多本已出版的卫生统计学教材，并将教学中积累的体会融入其中。希望该书能成为学生的良师益友。

在《卫生统计学教程》出版之际，首先向卫生统计学的前辈们致谢，感谢你们曾编写的卫生统计学书籍，使我们受益；同时感谢所有本教程中引用的文献的原作者。感谢陈育德教授在百忙中认真审阅了全部书稿，并提出了许多宝贵建议。感谢北京大学医学出版社为本教程出版提供了基金。

限于编者水平，本书难免有缺点错误，敬请专家和读者提出批评指正。

王 燕、康晓平

2006 年元月

# 目 录

<b>第一章 绪论</b> .....	(1)
第一节 学习卫生统计学的意义.....	(1)
第二节 卫生统计学中的一些基本概念.....	(2)
第三节 卫生统计工作的基本步骤.....	(5)
<b>第二章 数值变量的统计描述</b> .....	(8)
第一节 数值变量资料的频数分布.....	(8)
第二节 数值变量资料的描述指标 .....	(11)
第三节 正态分布 .....	(20)
第四节 正态分布的应用 .....	(23)
<b>第三章 分类变量的统计描述</b> .....	(27)
第一节 分类变量的频数分布及其统计指标 .....	(27)
第二节 标准化法及其应用 .....	(30)
第三节 动态数列及其应用 .....	(33)
<b>第四章 统计表与统计图</b> .....	(36)
第一节 统计表 .....	(36)
第二节 统计图 .....	(38)
<b>第五章 总体均数的估计和假设检验</b> .....	(50)
第一节 均数的抽样误差与标准误 .....	(50)
第二节 $t$ 分布 .....	(54)
第三节 总体均数的估计 .....	(55)
第四节 假设检验的一般步骤 .....	(58)
第五节 均数的 $t$ 检验和 $Z$ 检验 .....	(59)
第六节 均数的区间估计与假设检验的关系 .....	(68)
第七节 假设检验的两型错误和检验功效 .....	(70)
第八节 数值变量比较的秩和检验及数据变换 .....	(71)
第九节 假设检验中的其他若干问题 .....	(76)
<b>第六章 二项分布与 Poisson 分布及其应用</b> .....	(78)
第一节 二项分布的概念 .....	(78)
第二节 二项分布的应用 .....	(81)



---

第三节	Poisson 分布的概念	(83)
第四节	Poisson 分布的应用	(85)
<b>第七章</b>	<b>分类变量的假设检验</b>	<b>(88)</b>
第一节	成组设计四格表资料的 $\chi^2$ 检验	(88)
第二节	成组设计四格表资料的确切概率法	(92)
第三节	配对设计四格表资料的 $\chi^2$ 检验	(94)
第四节	$R \times C$ 列联表资料的 $\chi^2$ 检验	(95)
第五节	频数分布拟合优度的 $\chi^2$ 检验	(99)
第六节	单向有序分类变量的秩和检验	(100)
<b>第八章</b>	<b>实验设计</b>	<b>(103)</b>
第一节	医学实验设计概述	(103)
第二节	医学实验设计的要素	(104)
第三节	实验设计的基本原则	(105)
第四节	常用的几种设计方案	(110)
<b>第九章</b>	<b>调查研究与调查设计</b>	<b>(116)</b>
第一节	调查研究概述	(116)
第二节	调查设计	(119)
第三节	样本设计	(121)
第四节	调查技术简介	(128)
第五节	非抽样误差	(133)
<b>第十章</b>	<b>多组数值变量比较的假设检验</b>	<b>(136)</b>
第一节	方差分析的基本思想	(137)
第二节	单因素方差分析	(139)
第三节	均数之间的多重比较	(142)
第四节	协方差分析	(147)
第五节	Kruskal-Wallis 秩和检验	(152)
第六节	随机区组设计的方差分析	(155)
第七节	Friedman 秩和检验	(159)
<b>第十一章</b>	<b>简单线性相关与回归</b>	<b>(162)</b>
第一节	简单线性相关	(162)
第二节	秩相关	(166)
第三节	简单线性回归	(167)
第四节	线性回归的一些应用	(171)
第五节	非线性数据的线性转换	(174)

---

第六节 回归分析的注意事项	(176)
<b>第十二章 健康统计</b>	(177)
第一节 医学人口统计	(177)
第二节 疾病统计常用指标	(185)
<b>第十三章 寿命表及其应用</b>	(189)
第一节 寿命表的概念	(190)
第二节 寿命表的编制原理和方法	(191)
第三节 去死因寿命表	(198)
第四节 病例随访资料的生存分析	(202)
<b>参考书目</b>	(209)
<b>附录 1 统计用表</b>	(210)
<b>附录 2 SPSS for Windows 应用简介</b>	(233)
第一节 SPSS 软件包概述	(233)
第二节 现场调查数据的录入步骤	(233)
第三节 建立数据文件的 SPSS 命令	(234)
第四节 常见的数据预处理命令	(236)
第五节 实例操作	(239)
<b>附录 3 练习题</b>	(242)
<b>附录 4 汉英专业名词对照</b>	(266)

# 第一章 绪 论

## 第一节 学习卫生统计学的意义

统计学 (statistics), 是关于数据的收集、整理、分析、解释和表述的科学。统计学分成两个主要领域: 数理统计学和应用统计学。数理统计学侧重于建立统计方法和讲述统计方法的原理; 应用统计学则是结合特定专业研究特点, 使数理统计学原理与方法具体化, 从而产生加以前缀的统计学, 如, 社会统计学、心理统计学、生物统计学等。卫生统计学 (health statistics) 属于应用统计学的范畴, 是数理统计学的基本原理和方法在医学、特别是公共卫生学领域的应用, 是关于医学、特别是公共卫生研究中资料的收集、整理、分析、解释和表述的一门科学。卫生统计学是进行医学研究中认识事物数量特征与关系的一门方法学, 亦是为制定卫生政策提供定量依据的一门方法学。

为什么学习卫生统计学? 简言之为了进行医学领域的科学研究和科学决策, 现以公共卫生专业目前正在进行的两个科研项目为例加以说明。

**例 1.1** 世界卫生组织 (WHO) 在中国组织专家研发了一套关于儿童计划免疫的宣传材料及具体实施方法, 即儿童计划免疫的 IEC (information, education and communication) 策略, 在大规模推广之前, 需要在小范围实施并评价该 IEC 策略的有效性。北京大学公共卫生学院一个课题组承接了这个应用研究课题, 在最后能做出该 IEC 策略是否有效的结论之前, 我们需要决定: (1) 调查谁? (2) 调查几次? (3) 调查多少人? (4) 怎样得到调查对象? (5) 调查什么? (6) 如何调查? (7) 调查得到的资料如何整理? (8) 如何表示被调查者的儿童计划免疫知识、态度及行为的状况? (9) 如何分析儿童计划免疫 IEC 策略的有效性? 这一系列问题都是卫生统计学中欲解决的问题。

**例 1.2** 资料分析表明在全球范围内人类的乳腺癌、男性生殖系统癌症有明显上升的趋势; 同时, 男性的精子数量和浓度有明显减少的倾向。有学者提出了“环境激素”学说, 即人类越来越多地暴露于环境中有人体激素样作用、对人体内分泌起干扰作用的化学污染物质, 自此寻找“环境激素”的研究成为热点。北京大学公共卫生学院一个课题组欲研究人们经常食用的一种食品是否含有雌激素, 是否对小鼠有雌激素样作用。这个课题得到了国家自然科学基金的资助, 在进行课题申请书的撰写时, 必须进行统计学设计: 如 (1) 需要多少只大鼠? (2) 如何分组? (3) 如何设立对照? (4) 是否使用盲法? (5) 选用何种指标表示雌激素样作用? (6) 选用何种统计方法比较各组间指标的差别? 以及 (7) 如何解释实验结果等。上述问题的解决都离不开卫生统计学, 大家在学习过程中会逐步找寻到答案。

在学习卫生统计方法之前, 我们首先需要了解卫生统计学中的一些基本概念和卫生统计工作的基本步骤。

## 第二节 卫生统计学中的一些基本概念

### 一、观察单位 (observation unit) 与变量 (variable)

观察单位是指被观察或测量对象的最基本单位, 亦称个体, 可以是一个人、一只鼠、一个样品、一个采样点或一个地区等。对每个观察单位的某项特征进行测量或观察, 该项特征称为变量, 得到的被观察单位的该项特征值称为变量值 (value of variable), 亦称观察值或测量值。例如, 例 1.1 的研究其中一项内容是了解某地区 2 岁以下儿童的卡介苗接种情况, 课题组检查了该地区 200 名 2 岁以下儿童的卡疤, 这个例子中观察单位为一名 2 岁以下儿童, 变量为卡疤, 变量值为“+”或“-”。

### 二、变量的类型

变量以其变量值的特点, 分为两大类, 即数值变量和分类变量, 分类变量又可分为无序分类变量和有序分类变量。不同类型的变量需要选用不同的统计指标和统计方法进行分析。根据分析需要, 不同类型变量之间可进行转换。

#### (一) 数值变量 (numerical variable)

通过测定每个观察单位的某项特征的大小所得到的数据, 称为数值变量, 其变量值是以数值表示的, 通常有度量衡单位。例如, 调查某地 2 岁男孩的身长发育状况, 这时一个 2 岁男孩是观察单位, 测量指标, 如身高 (cm)、体重 (kg)、血红蛋白 (g/L)、牙齿数 (个) 就是数值变量。描述数值变量常用的统计指标有平均数, 标准差等 (见第二章)。统计分析方法有  $t$  检验,  $Z$  检验, 直线相关与回归, 方差分析等 (见第五、十、十一章)。

#### (二) 分类变量 (categorical variable)

通过确定每个观察单位的某项特征的性质或类别得到的数据, 称为分类变量, 其变量值是定性的, 表现为互不相容的类别或属性, 没有度量衡单位。例如, 例 1.1 的研究得到的每个儿童卡疤“+”或“-”的数据就是分类变量。通常, 作为对分类变量资料进行初步整理, 先按类别将观察单位分组, 如分为“+”组和“-”组, 然后清点每组中的人数, 这样得到的数据称为计数资料。描述分类变量常用的统计指标有比率、速率等 (见第三章), 统计分析方法有  $Z$  检验,  $\chi^2$  检验 (见第六、七章)。

分类变量又可分为几种类型:

1. 无序分类变量。包括①二项分类变量, 特点是其变量值分为两类, 如, 检查 2 岁儿童卡疤得到的阳性或阴性; 观察某药对某病患者疗效得到的有效或无效。②多项分类变量, 特点是其变量值分为两类以上, 如, 职业、血型等变量。

2. 有序分类变量。特点是其变量值是多项分类且各类之间有程度的差别。如, 文化程度, 可分为: 没上过学、小学、初中、高中和大专及以上等; 疗效, 可分为治愈、显效、有效和无效。针对这类变量的统计分析方法有秩和检验和等级相关分析等。

#### (三) 数据转换 (data transformation)

根据分析的需要, 数值变量可转换为分类变量。例如, 观察得到 100 名婴儿出生体重 (克), 这是数值变量资料。如果欲分析低出生体重婴儿所占的比例, 可以将出生体重 < 2500

克定义为低出生体重； $\geq 2500$ 克定义为非低出生体重两类，这时就成了二项分类变量。如果分组再细一些，将出生体重 $< 2500$ 克定义为低出生体重， $2500 \sim 3999$ 克定义为正常出生体重， $\geq 4000$ 克定义为高出生体重，这时数值变量就成了有序分类变量。

很多情况下，为了便于计算机的识别和运算，对分类变量可以进行赋值。例如，男女分别赋值为1和2，文化程度按文盲、小学、初中、高中、大专及以上分组，可分别赋值为0、1、2、3、4、5。这种赋值仅是一种“数据代码”，这些变量的本质还是分类变量，应该按分类变量进行统计分析。

### 三、同质 (homogeneity) 与变异 (variation)

研究对象具有的相同的状况或属性等共性称同质或同质性。对于同质的各观察单位，其某变量值之间的差异，称为变异，例如，研究某地2005年活产婴儿的出生体重，同质是指同一地区、同一年份、同为活产；这些活产婴儿出生体重不尽相同，存在差异，这种体重值之间的差异就是变异。又如，研究某新药治疗胃溃疡的效果，所有研究对象都必须是确诊为胃溃疡的病人且病情相似，在这种同质的基础上观察治疗效果，有的人治愈，有的人未愈，这种差异就是变异。卫生统计学所研究的对象都是以同质为基础，并具有变异的事物，这也是之所以用变量这一术语来表示观察单位的特征的理由。

### 四、总体 (population) 与样本 (sample)

总体是根据研究目的确定的同质观察单位的全体，确切地说，是同质的所有观察单位某种变量值的集合。例如，欲研究某地2005年活产婴儿的出生体重，该地2005年所有活产婴儿的出生体重值就构成一个总体。又如，例1.1的研究，欲了解某地区2005年4月20日2岁以下儿童的卡介苗接种情况，该地该时所有2岁以下儿童的卡疤情况就构成一个总体。这两个例子的总体都明确了一定时间、一定空间，理论上说观察单位的数量是可知的、有限的，称为有限总体。有时总体是抽象的，如欲研究某药治疗胃溃疡的效果，这里总体是指所有胃溃疡病人，但没有时间和地点的限制，观察单位总数量是不可知的，该总体称为无限总体。

样本是指总体中的一部分观察单位的某项变量值的集合，这一部分必须是对总体具有代表性的，或说是总体的缩影。对于有限总体，保证样本对总体具有代表性的手段是随机抽样 (random sampling)，所谓随机抽样就是总体中每个观察单位都有均等的机会被抽中作为样本，抽中哪个作为样本具有一定的偶然性，具体的随机抽样方法见第九章。例如，欲调查某地2005年活产婴儿的出生体重，从该地区2005年出生婴儿随机抽取200名，测量其出生体重，这200名婴儿的出生体重值就是样本。对于无限总体，只有通过明确样本定义，从而抽象出样本是某总体的缩影，或说样本对某总体具有代表性，如用某药治疗了200例胃溃疡病人，它的总体就是所有服用该药的胃溃疡病人。

通常，医学研究不可能也没必要对总体中的每个观察单位进行观测或检测，例如，确定某品牌冰棍是否符合卫生标准，只能抽取一定的样本进行检测；再如，欲研究某药治疗胃溃疡的效果，也只能治疗部分病人。通常情况下，医学研究是对样本进行研究，也称为抽样研究 (sampling study)，但其目的是通过样本的信息去推论总体的特征。如何用样本信息推论总体，正是卫生统计学的奥妙所在。

## 五、误差 (error)

任何周密设计的科学研究,都不可能没有误差。医学研究中的误差通常指测量值与真值之差,其中包括系统误差、随机测量误差和抽样误差。抽样误差是样本统计指标与总体参数(总体统计指标)之差,是统计学研究和处理的重要内容。随机测量误差及抽样误差又同属于随机误差。

### (一) 系统误差 (systematic error)

系统误差是某种必然因素所致,不是偶然机遇造成的,具有一定的方向性,使观察结果一律偏高或偏低。系统误差一旦发生,统计学是无能为力的,因此要尽可能避免。大多数系统误差可以通过周密的研究设计和调查(或测量)过程中的严格质量控制措施得以解决。系统误差发生的常见情况包括:①操作方法不正确或问卷调查时方法有误;②医生掌握疗效标准偏高或偏低;③周围环境的改变,如实验室内室温过高或过低以及现场调查时出现不必要的行政干预;④仪器不准或试剂不合格,例如,测量血压要求血压计的水银面与0平行,如果使用的血压计没校正,高出4 mmHg,那么测定出的血压值都高4 mmHg。

### (二) 随机测量误差 (error of random measurement)

随机测量误差是偶然机遇所致,故无方向性,对同一样品多次测定,结果有高有低,不完全一致。随机测量误差是不可避免的,再精确的测量仪器也会存在误差,但只要将误差控制在一定的允许范围内,读出的数据都可以使用。

### (三) 抽样误差 (sampling error)

在抽样研究中,即使消除了系统误差,控制了随机测量误差,样本统计指标与总体参数间仍会存在差别,称这种差别为抽样误差。抽样误差是由于个体差异造成的,是抽样所致,是客观存在、不可避免的。抽样误差可以通过统计方法进行估计,也可通过增大样本使其减小。我们可以通过一个实验来理解什么是抽样误差。假定已知某年某地所有1000名13岁女学生身高的总体均数( $\mu$ )是155.4cm,该地每一个13岁女学生都有一个身高测量值,我们将这1000名女生的身高值(cm)都录入计算机,存在数据库里作为一个有限总体。在这样一个有限的总体中做多次重复抽样,每次均抽取30例( $n_i=30$ )组成一个样本,可以算出每一个样本的平均身高( $\bar{X}_i$ ),因为是完全随机抽样,数据库中的每一个学生的身高值都有可能被抽到,最终得到的样本均数( $\bar{X}_i$ )可能是153.6,153.1,154.9...158.7等。可以看到样本均数( $\bar{X}_i$ )与总体均数( $\mu$ )间有一个差别,而且样本均数与样本均数间也有差别,这种误差既不是系统误差,也不是测量误差,完全是由抽样造成。因此我们讲,只要是对存在变异的观察单位进行抽样研究,必然存在抽样误差,这种误差虽然是不可避免的,但可以认识它、估计它并可缩小它。

## 六、概率 (probability) 与频率 (frequency)

概率与频率都是表示某事件发生的可能性大小的数值。概率是对总体而言,频率是对样本而言。概率用符号 $P$ 来表示,数值在0与1之间,即 $0 \leq P \leq 1$ ,也可用百分数表示。 $P$ 越接近1,表明某事件发生的可能性越大, $P$ 越接近0,表明某事件发生的可能性越小。频率可用小写 $p$ 表示,取值范围及意义与概率相同。如某药治疗200个病人,其治愈率为80%,这80%是频率。频率是从一次试验或一个样本计算得到的某事件发生率,若经过多次试验

或许多人的治疗，其治愈率稳定在 80%，这时可以说，某药治愈某病的可能性，即概率为 80%。卫生统计学中的许多结论都是根据概率得到的。一般常将  $P \leq 0.05$  或  $P \leq 0.01$  称为小概率事件，表示某事件发生的可能性很小、是不可能发生的事件，具体的应用在以后的章节中将会介绍。

### 第三节 卫生统计工作的基本步骤

设计、收集资料、整理资料、分析资料与解释结果是卫生统计工作的四个基本步骤，这四个步骤是紧密联系不可分割的，某一环节发生错误，都可影响研究结果的正确性。

#### 一、设计 (design)

设计是开展研究工作的前提和依据，一个完整的设计应包括研究全过程的内容，具体包括研究意义、研究目的、研究假设、研究内容、研究方法、研究对象、抽样方法、样本含量、问卷设计、统计指标、分析方法、资料整理、质量控制、预期结果、经费预算、人员安排和进度等等。卫生统计学要解决的设计主要是统计学设计，主要是围绕资料收集、整理、分析这一过程的设计，具体包括针对研究方法、研究对象、抽样方法、样本含量、问卷设计、统计指标、分析方法、整理资料方法以及质量控制方法等方面的设计。

#### 二、收集资料 (collection of data)

收集资料的任务是按照设计要求取得准确可靠的原始数据。

##### (一) 卫生统计资料的来源

卫生统计资料的来源是多方面的，可概括为经常性资料和一时性资料两大类：

1. 经常性资料：一般指医疗卫生工作中的记录。①统计报表，如医院工作报表、居民病伤死亡原因报表、疫情报表、妇幼卫生年报表等。②医疗卫生工作记录和报告单（卡），如医院病历、健康检查记录、各种医学检验记录及传染病报告卡等。

2. 一时性资料：为某项研究而专门设计的现场调查、实验或试验。

##### (二) 卫生统计资料的要求

原始资料是卫生统计工作的基本依据，俗话说“烂棉花织不出好布”，把好收集资料质量这一关非常关键，要努力做到：

1. 资料完整、正确。完整是指调查项目填写完整无空项，若确实不详可用代码填写，如年龄不详，可填“99”。正确是指填写的内容准确无错误。

2. 有足够的数量。原始数据要有一定的数量才能反映事物的规律性，但并不是越多越好，足够即可。多少数量达到足够本书第八章将讲授。

3. 具有代表性、可比性。代表性是指样本对总体要有代表性。对于有限总体，随机抽样保证样本的代表性；对于无限总体，明确样本的定义，可推测样本代表的总体。可比性是指两组或多组资料比较时，除观察问题或实验因素不同外，其他因素要求尽量一致，例如，比较两种药物治疗胃溃疡的疗效，两组病人除了用药不同外，其他因素，如病情等，应尽可能一致。保证可比性的方法是随机化分配（见第八章）。

### 三、整理资料 (sorting of data)

整理资料的任务是清理原始数据,使其条理化、计算机数据化,以便进一步计算指标和分析。

#### (一) 原始数据的检查与核对

原始数据的常规检查包括:①检查原始记录的数据有无错误和遗漏;②各项目是否按要求或填表说明填写;③有无不合逻辑的项目;④统计报表的行栏合计是否与总计相符等。这部分检查核对应在调查现场时做,以便及时更正。

#### (二) 建立数据库、进一步净化数据

1. 利用计算机数据库软件建立数据库。

2. 将原始数据录入计算机。最好采取双人录入方法以避免录入过程的错误,双人录入方法,即两个人分别录入一份原始数据,然后核查并更正录入不一致的数据。

3. 数据的取值范围检错。可以在数据库建立时,对某些变量的取值范围给予规定,如性别,“1”、“2”或“男”、“女”;出生体重,1500~6000g等;也可利用频数分布表检查是否有异常值的出现,如在“结婚年龄”的频数表中出现“15岁”的,这时要与原始数据核对。

4. 数据间的逻辑关系检错。逻辑检查是为了检查变量值之间是否有矛盾。例如,吸烟的调查,某一被调查者的年龄填写“23岁”,吸烟史填写“20年”,这意味此人3岁就开始吸烟,显然是不可能。数据间的逻辑关系检错可以通过编写计算机语句、利用计算机完成。

对于通过范围检错和逻辑关系检错,检查出来的不合理的数据,尽量通过重新调查进行更正,如果不可能重新调查,只能在分析时剔除不合理的数据或列为不详。

### 四、分析资料 (analysis of data) 与解释结果

分析资料的任务是按研究设计的要求,结合变量的类型计算有关指标,阐明事物的内在联系和规律。统计分析主要包括:①用一些统计指标、统计图表等描述资料的数量特征和分布规律。②对样本统计指标做参数估计和假设检验,目的是用样本信息推论总体特征。最后,还需要结合卫生统计学知识与专业知识对分析结果做出恰当的解释。

分析资料的方法在本书中占有大部分篇幅,是本书的主要内容,但绝不能认为分析资料是卫生统计工作的全部,卫生统计工作的四个基本步骤,即设计、收集资料、整理资料、分析资料与解释结果是紧密联系不可分割的,某一环节发生错误,都可影响研究结果的正确性。

另外,由于计算机的普及,除了设计和资料收集需要大量的人工操作外,整理资料和分析资料都可在计算机上完成。一些统计软件,例如,SPSS、SAS、Stata等都可以做数据的录入、检错和分析。用计算机替代人工操作,确实提高了效率,但“电脑”是不能完全替代“人脑”的。例如,如何分组?分几组?计算什么指标?用什么图表示?选用什么方法进行假设检验?对分析结果的解释等,都是“人脑”决定的,计算机仅是操作的工具。有了计算机及计算机软件这一强有力的工具,并不意味着人工分析计算就没有必要了,通过人工分析计算,可以帮助我们加深对分析指标和方法的理解、记忆和运用。本书在重点讲授人工分析操



作的基础上,介绍了 SPSS 软件的操作。对学生的要求是以掌握卫生统计学的基本概念、基本思路、基本方法和人工分析计算为主,计算机 SPSS 软件的操作作为一个基本技能,有助于学生今后的学习与工作。

(王 燕、康晓平)