卫生部规划教材

全国中等卫生学校教材

供预防医学专业用

卫生统计学

第三版

主编 周士楷

人民卫生出版社

图书在版编目（CIP）数据

卫生统计学/周士楷主编；—北京：人民卫生出版社，1996
ISBN 7-117-02401-1

Ⅰ．卫… Ⅱ．周… Ⅲ．卫生学：统计学
Ⅳ．R195.1

中国版本图书馆 CIP 数据核字（96）第 16734 号

卫生统计学

第 三 版

周士楷 主编

人民卫生出版社出版
（北京市崇文区天坛西里10号）
三河市宏达印刷厂印刷
新华书店北京发行所发行

787×1092 16 开本 13 $\frac{1}{4}$ 印张 302 千字
1987年5月第1版 1997年4月第3版第12次印刷
印数：159,151—184,150
ISBN 7-117-02401-1/R·2402 定价：11.10元

第三轮中等医学教材出版说明

卫生部曾于1983年组织编写、陆续出版全国中等卫生学校11个专业使用的77种教材。1992年又组织小修订，出版第二轮教材。为我国的中等医学教育作出了积极贡献。

为适应中等医学教育改革形势的需要和医学模式的转变，1993年11月，卫生部审定、颁发了全国中等卫生学校新的教学计划及教学大纲。在卫生部科教司领导下，我们组织编写（修订）出版第三轮全国中等医学12个专业96种规划教材，供各地教学使用。

这轮教材以培养中级实用型卫技人才为目标，以新的教学计划及大纲为依据，体现“思想性、科学性、先进性、启发性、适用性”，强调“基本理论知识、基本实践技能、基本态度方法”。教材所用的医学名词、药物、检验项目、计量单位，注意规范化，符合国家要求。

编写教材仍实行主编负责制；编审委员会在教材编审及组织管理中，起参谋、助手、纽带作用；部分初版教材和新任主编，请主审协助质量把关。第三轮中等医学教材由人民卫生、河北教育、山东科技、江苏科技、浙江科技、安徽科技、广东科技、四川科技和陕西科技九家出版社出版。

希望各校师生在使用规划教材的过程中，提出宝贵意见，以便教材质量能不断提高。

卫生部教材办公室

1995年10月

全国中等医学教材编审委员会

第三版前言

本书第三版是根据1994年12月全国中等医学教材会议精神和新颁教学计划、大纲的要求进行编写的，供中等卫生学校预防医学专业用。

全书分十五章，第一章绪论，根据大纲要求，加强了其中基本概念部分的论述；第二章至第十四章讲解基本统计方法，根据目标教学的要求，本书加强了分类资料统计分析方法的介绍，删除了原教材中简捷计算法、多元分析数学演算等内容；第十五章为居民健康统计，新增了寿命表的应用。

教材编写会议期间，经本专业各学科主编研究协调，对大纲内容作少许调整：大纲中半数致死量一章，由《卫生毒理学基础》编写，本书不再保留。课内实习安排10学时，用于必须在课堂内完成的作业。为加强统计思维和资料分析能力的培养，将复习思考和作业题附于各章正文之后，供学生课外及时复习思考和练习用，不再在课内安排做练习题之类的“实习课”。

本书中有“*”号的内容供选学用。书中例题及练习题的数据引自参考书、各专业杂志或内部资料。书中插图由陈嘉冰老师绘制，在此一并向有关作者表示衷心的感谢。书中所有统计资料仅供演算示范或数据处理练习用，不作为有关专业内容的解释依据。书中统计符号、缩写一律按《中国医学百科全书·医学统计学分卷》的规定使用。

限于编者的水平和经验，书中一定存在着缺点和错误，希望使用本书的师生提出批评指正，以利改正。

编　者

1996年2月

目　录

第一章　绪论 …… 1
第一节　概述 …… 1
第二节　卫生统计工作的步骤 …… 2
第三节　几个基本概念 …… 3
第二章　数值资料的统计描述 …… 8
第一节　数值资料的频数分布 …… 8
第二节　集中趋势指标 …… 10
第三节　离散程度指标 …… 15
第四节　正态分布 …… 19
第三章　总体均数的估计和 t 检验 …… 24
第一节　均数的抽样误差 …… 24
第二节　t 分布 …… 24
第三节　总体均数可信区间的估计 …… 26
第四节　t 检验 …… 28
第五节　进行 t 检验时应注意的问题 …… 33
第四章　方差分析 …… 39
第一节　完全随机设计资料的方差分析 …… 39
第二节　随机区组（配伍组）设计资料的方差分析 …… 44
第三节　方差分析的基本思想 …… 46
第四节　多个均数间的两两比较 …… 47
第五节　方差分析的基本条件与数据变换 …… 49
第五章　分类资料的统计描述 …… 54
第一节　分类资料的频数表 …… 54
第二节　常用相对数 …… 54
第三节　动态数列 …… 56
第四节　应用相对数时应注意的问题 …… 58
第五节　标准化法 …… 59
第六章　二项分布及其应用 …… 63
第一节　二项分布的概念及应用条件 …… 63
第二节　二项分布特征 …… 64
第三节　二项分布的应用 …… 66
第七章　泊松分布及其应用 …… 70
第一节　泊松分布的概念 …… 70
第二节　泊松分布的性质 …… 71
第三节　泊松分布的应用 …… 72

第八章　χ^2　（卡方）检验 …… 79
第一节　样本率或构成比的比较 …… 79
第二节　分类资料分层分析——MHχ^2 检验 …… 85
第三节　分类资料的相关分析 …… 87
第四节　频数分布拟合优度检验 …… 89
第五节　四格表资料的确切概率法 …… 90
第九章　非参数统计 …… 95
第一节　非参数统计的概念 …… 95
第二节　秩和检验 …… 95
第十章　直线相关与回归 …… 103
第一节　直线相关 …… 103
第二节　等级相关 …… 106
第三节　直线回归 …… 107
第四节　作相关与回归分析时应注意的问题 …… 112
第十一章　多元线性回归简介 …… 116
第一节　多元线性回归的概念 …… 116
第二节　多元线性回归分析实例 …… 116
第三节　多元线性回归分析的若干指标 …… 118
第四节　多元线性回归的应用 …… 119
第十二章　统计表与统计图 …… 122
第一节　统计表 …… 122
第二节　统计图 …… 125
第十三章　调查设计 …… 130
第一节　调查设计的意义 …… 130
第二节　搜集资料的计划 …… 130
第三节　整理资料的计划 …… 134
第四节　分析资料的计划 …… 136
第五节　样本含量的估计 …… 137
第六节　混杂因素干扰的排除 …… 138
第十四章　实验设计 …… 140
第一节　实验研究的基本要素及实验设计的意义 …… 140
第二节　实验设计的基本原则 …… 141
第三节　实验设计方法 …… 142
第四节　样本含量的估计 …… 144
第十五章　居民健康统计 …… 147
第一节　生育与计划生育统计 …… 147
第二节　人口死亡统计 …… 149
第三节　简略寿命表 …… 152
第四节　疾病统计 …… 155

附录…… 160
附录Ⅰ 矩法正态性检验…… 160
附录Ⅱ 统计分析软件包在数据处理中的应用…… 162
附表…… 166
附表1 标准正态曲线下的面积表 …… 166
附表2 t 界值表 …… 168
附表3 F 界值表（方差齐性检验用）…… 169
附表4 F 界值表（方差分析用）…… 170
附表5 q 界值表 …… 176
附表6.1 Dunnett t 表（单侧检验） …… 177
附表6.2 Dunnett t 表（双侧检验） …… 178
附表7 百分率与概率单位对照表 …… 179
附表8.1 百分率的可信区间（$1 \leqslant n \leqslant 50$） …… 181
附表8.2 百分率的可信限（$50 \leqslant n \leqslant 100$） …… 184
附表9 泊松分布均数的可信限 …… 185
附表10 χ^2 界值表 …… 186
附表11 阶乘的对数表 …… 187
附表12 相关系数 r 界值表 …… 188
附表13 等级相关系数 r 界值表 …… 190
附表14 配对比较的秩和检验 T 界值表 …… 191
附表15 两组比较的秩和检验 T 界值表 …… 192
附表16 三组比较的秩和检验 H 界值表 …… 194
附表17 M 界值表 …… 194
附表18.1 两样本率比较时所需样本含量（单侧） …… 195
附表18.2 两样本率比较时所需样本含量（双侧） …… 196
附表19 配对比较（t 检验）时所需样本含量 …… 197
附表20 两样本均数比较（t 检验）时所需样本含量 …… 198
附表21 随机数字表 …… 199
附表22 10×10 随机化拉丁方 …… 200
附表23 30个自然数的随机排列表 …… 201
参考书目…… 202

第一章 绪 论

第一节 概 述

统计学是研究数据的搜集、整理、分析与推断的科学。卫生统计学是把统计理论与方法，应用于居民健康状况、医疗卫生工作实践和医学科学研究的一门学科。现代医学的研究大量采用定量研究的方法，往往通过调查或实验取得数据，进而从数据中提取有用的信息。由于错综复杂的条件和难以控制的因素的影响，医学研究中所获得的数据，一般变异较大，不能直接从中了解事物本来的面貌和其中所蕴藏的规律。卫生统计学的一个任务是借助于统计方法，从有限的观察及表现为偶然的数据中，把所研究的事物或现象的本质特征、整体情况和与其他事物或现象间的关系一一揭示出来。

（一）卫生统计学的主要内容

1. 卫生统计学的基本理论和方法　讨论统计设计、统计描述、统计推断的原理和方法。研究统计方法在医学科学，尤其是在预防医学和卫生管理学中的应用。

2. 健康统计　包括出生、死亡、疾病、身体发育统计等。其任务是搜集有关资料，建立和应用有关指标，反映和评价居民健康状况，探讨并掌握疾病在人群中的发生、发展、流行、分布的特点与规律，为疾病防治和卫生保健计划与决策提供依据。

3. 卫生事业统计　搜集、整理和分析卫生事业发展的情报（如卫生机构、人员、设备等卫生资源及其利用情况），为更有效地开发利用卫生人力、物力和财力资源，为卫生事业的宏观管理和计划决策提供科学的依据。

本教材将以较大篇幅介绍卫生统计的基本理论和方法，同时重点介绍健康统计的部分内容。

（二）卫生统计学的应用和发展

卫生统计学在医学研究、疾病防治和卫生事业管理方面有着广泛的应用。例如：在预防医学工作中，为了了解居民健康状况和评价医疗卫生措施的效果；为了判断社会-心理、环境和生物因素对居民健康的影响；为了改善卫生条件，提高居民健康水平，经常需要作流行病学、卫生学调查和实验研究。卫生统计学是完成这些任务的重要手段之一，因此从事预防医学的工作者，必须学好卫生统计学。

应用数学的发展，为卫生统计学的发展提供了理论基础；预防医学的发展不断向卫生统计学提出了新的要求；大量的卫生统计工作实践，积累了丰富的经验，进一步充实了卫生统计学的内容和扩大了卫生统计学的应用范围。近年来，由于电子计算机的普遍应用，资料搜集、整理、储存、检索、分析、传输与交流工作已经实现了自动化，使人们摆脱了过去那种繁琐的手工处理数据的劳动。过去因计算繁琐而难以采用的多元统计分析方法，目前已经成为实际工作中广为流行的统计分析方法。由于建立了统计资料的数据库，实现了数据共享，使统计资料中所蕴藏的丰富的信息得以被充分地提取和利用。

（三）学习卫生统计学的目的与要求

预防医学专业学生学习卫生统计学的目的是：掌握卫生统计学的基本理论、基本知

识和基本技能、为学习各门专业课、阅读专业书刊、从事预防医学工作打下必要的统计学基础。为此要求：

1. 明确学习目的　充分认识卫生统计学在医疗卫生工作中的重要作用，发挥学习的主动性和积极性。

2. 注意学习方法　要弄清统计方法的基本思想及有关公式的应用条件和用法，但不深究其数学原理；要理论联系实际，统计设计和分析都要结合专业知识，但不作专业上无意义甚至错误的统计结论；要重视基本理论与知识的学习，把各章节的内容作为一个整体，前后呼应，纵横联系，融会贯通地学习。

3. 培养科学作风　通过学习，培养对数据处理的严肃认真、实事求是的科学作风，重视原始资料的搜集与整理的每一个细节，保证资料真实可靠。

4. 培养科学的统计思维方法　养成依据统计学原理思考问题，进行分析、判断和推理的习惯。这里所指的统计学原理主要有：变异的客观存在；抽样误差及其控制；混杂因素干扰及其排除；假设检验的基本思想及其结论的概率性等。

第二节　卫生统计工作的步骤

统计全过程设计、搜集资料、整理资料和分析资料是卫生统计工作的四个基本步骤。

1. 统计全过程设计　根据研究目的，从统计学角度，对搜集资料、整理资料和分析资料提出周密的计划和要求，作为统计全过程实施的依据，以便能用尽可能少的人力、物力和时间，获得准确可靠的结论。

统计全过程设计要求科学、周密、简明。关于这方面的内容将在第十三章和第十四章详细介绍。

2. 搜集资料　根据资料的来源，可将资料分为经常性和一时性资料两种。经常性资料包括从日常医疗卫生工作原始记录（如病历）、专门报告卡（如出生、死亡报告卡）、统计报表（如疫情月报表、年报表）中搜集到的资料。一时性资料常指由现场调查和实验室研究搜集的资料。现场调查一般涉及范围较广，观察数量较多，影响因素较复杂；实验研究一般涉及范围较小，观察数不多，因而便于对实验条件作严格控制，结论也比较确切。现场调查与实验研究既有区别又有联系，它们常结合使用。现场调查可提供进一步实验研究的线索，而实验研究的成果也需要再回到现场中去实践验证。

搜集资料要求完整、准确、及时。完整是指搜集资料项目不遗漏；准确是指观察、测量准确，记录、计算无误；及时是指经常性资料的搜集，应按规定的时间完成，一时性资料的数据记录应在观察、测量的同时完成，不得以“回忆”方式记录数据。

3. 整理资料　把搜集到的资料，有目的、有计划地进行科学加工，使分散的、零乱的资料变成系统化、条理化的资料，以便进一步分析。

4. 分析资料　运用各种统计分析方法，结合专业知识，从经过整理的资料中，计算出各种指标，然后对事物或现象进行统计描述与统计推断，揭示其本质特征、整体情况和与其他事物或现象间的关系。要做好分析资料，应具备两个条件：①对于各种统计分析方法能够融会贯通地理解，能够正确地选择、综合地运用各种统计分析方法；②对于所研究的事物本身及其与周围事物的联系具有丰富的知识，因而能够作出合理的判断。

第三节　几个基本概念

一、观察单位

观察单位（或个体）是获得数据的最基本的、最小的单位。观察单位可以是人、标本、家庭、国家等。例如，欲了解健康男子的血红蛋白量，一名健康男子就是一个观察单位；欲了解家庭的年总收入，则一个家庭就是一个观察单位；欲了解中专学校在校学生数，则一所中专学校就是一个观察单位。

二、变　　异

世间事物或现象普遍地存在着变异性。在医学研究中，有两种变异。一种称为个体变异，即观察单位本身的变异，表现为即使各个条件相同的个体，其某项特征仍存在着差异。例如，同一年龄的男孩，身高、体重各不相同；同一年龄的成年男子，血清中胆固醇含量各不相同；同一种病的患者，即使病情一致，治疗方法和条件相同，疗效也未必相同等等。另一种称为随机测量变异，它是由于测量手段或条件的波动而造成测量结果的差异。例如，同一样品，用分析天平多次称量，各次所得量值不完全相同。

三、变量和变量值

1. 变量　变量是观察单位的某项特征（或某种属性），如上述健康男子的血红蛋白量、家庭的年总收入、中专学校的在校学生数等都是变量；学生的性别、籍贯也是变量。把观察单位的某项特征称为变量，是因为在众多的观察单位中，某项特征的观察结果可以各不相同。如健康男子血红蛋白量可以不同，家庭的年总收入可以不同，中专学校在校学生数可以不同，学生的性别、籍贯也可以不同等。

2. 变量值　又称观察值，是指对于观察单位的某项特征（变量）的观察结果（包括分类鉴别、程度测量、量值测定等）。

四、变 量 类 型

按观察值的性质不同可以把变量分为数值变量和分类变量：

1. 数值变量　又称定量变量，其变量值一般为带有度量衡单位的大小不等的数值，可分为连续性和非连续性两种。如身高、体重等为连续性数值变量，某医院每月门诊人数、育龄妇女生育人数等为非连续性数值变量。

2. 分类变量　又称定性变量或字符变量。其变量值是代表互不相容类别或属性的字符。例如，性别、民族，其变量值可以是男或女、汉族或其他各民族。分类变量包括两项分类变量和多项分类变量，如性别分为男女两项互不相容的类别，民族分为汉、藏、维吾尔、苗等多项互不相容的类别。

分类变量又分为无序分类变量和有序分类变量。无序分类变量指变量值间并无程度或等级的差别，如前述性别、民族就是二项无序分类变量和多项无序分类变量。如果分类变量值之间，存在着程度或等级上的差别，这种程度或等级带有“半定量”的性质，这样的变量就称为有序分类变量或等级变量。如检查尿蛋白，观察单位为每一份尿标本，观

察结果可以分为－、＋、＋＋、＋＋＋、＋＋＋＋等各类（尿蛋白量一类比一类多），又如观察某病治疗效果，观察结果可分为痊愈、显效、好转、无效、恶化、死亡等类（疗效一类比一类差）。

以数值变量值为原始数据的统计资料称数值资料（定量或计量资料）。以分类变量值为原始数据，清点并汇总具有不同类别变量值的观察单位的个数，编制成分类变量频数表的统计资料称分类资料（定性或计数资料），其中以等级变量值为原始数据，分类计数其观察单位个数，编制成频数表的资料称等级资料（或半定量资料）。

数值资料、分类资料或等级资料，各有其适用的统计分析方法。分析者首先要鉴别资料是属于何种类型，然后选择适用的统计分析方法进行分析。但资料类型并不是绝对的，根据研究目的或分析需要，可将数值资料转换为分类资料，也可把分类资料转换为数值资料。如白细胞总数资料为数值资料，也可按白细胞总数正常或不正常分为两类，清点并汇总各类观察单位的个数，此时即为分类资料。若将白细胞增加的情况区分为轻度增加、中度增加、重度增加三类，清点并汇总其观察单位个数，就成为有序分类资料，即等级资料。有时需要将分类变量值或等级变量值作“数量化”处理，使之转换为0，1，2，3……数值，这种经过“数量化”的分类资料或等级资料可作为数值资料进行分析。

五、总体与样本

总体是根据研究目的确定的性质相同的所有观察单位某种变量值的集合。例如，欲研究某地40岁以上正常男子血清胆固醇含量，则该地所有40岁以上的正常男子的血清胆固醇量值的集合就是总体。在这里，性质相同的观察单位指的是该地的（不包括外地的）、所有的（不是部分的）、40岁以上的（不包括40岁以下的）、正常的（不包括有病的，这里的“病”是指某些影响胆固醇含量的疾病）男子（不包括女子）。

了解总体情况是统计工作的目的之一。但是，多数情况下，总体很大，我们不可能对所有观察单位逐个观察。有时由于观测带有破坏性，如检查一批鸡蛋中沙门菌污染情况，即使可能逐一检查，也不允许这样做。因此，常常是通过样本情况去推断总体情况。

样本是从总体内随机抽取的一部分。例如，从某地抽取100名40岁以上的正常男子，测定其血清中胆固醇含量，这100个胆固醇值就组成了一个样本，为了使样本对总体具有充分的代表性，必须遵循随机抽样原则，即要使总体内每一个体都具有同等被抽取到的机会。用随机抽样方法建立的样本称随机样本，本书介绍的许多由样本信息推断总体特征的方法，都是建立在随机样本的基础上的，如果样本不是采用随机方法建立，则使用某些统计方法就失去了理论依据。关于随机抽样方法见第十三章。

六、概率与频率

概率是事件发生的可能性大小的度量，以符号P表示。可以把我们经常遇到的事件分为三种类型：①必然事件：指的是必然会发生的事件。如气压为101.325千帕（1个标准大气压）时，纯水加热到100℃，必然会发生沸腾现象。必然事件的概率$P=1$。②不可能事件：指的是不可能发生的事件。如从地球上看，太阳从西边升起的事件，必定不会发生。不可能事件的概率$P=0$。③随机事件：指的是在一定条件下，可能发生，也可能不发生的事件。如某人在肝炎流行时，是否会得肝炎，回答是不能肯定的，可能得，也

可能不得。随机事件的概率 P 在 0 与 1 之间，某事件发生的可能性愈大，则其概率 P 愈接近 1；某事件发生的可能性愈小，则其概率 P 愈接近 0。在统计学中有一个公认的道理，即“小概率事件在一次观察中，可以认为不会发生”。

频率也是某事件出现的可能性大小的度量，只不过概率是对总体而言，频率是对样本而言。在相同的条件下，进行 n 次重复试验，事件 A 发生数为 $a(a \leqslant n)$，则 a 与 n 的比为事件 A 的频率。如 n 逐渐增大，则事件 A 的频率就越来越接近事件 A 的概率 P，因此，常以 n 充分大时事件 A 的频率作为该事件概率的近似值。

七、误　　差

统计上所说的误差，包括测得值（观察值）与真值之差和样本指标与总体指标之差。从误差的性质来看，可以把误差分为两大类，即偶然误差和系统误差。

（一）偶然误差

偶然误差又称随机误差，它包括抽样误差和随机测量误差。

1. 抽样误差　指样本指标（如样本均数、样本率）与总体指标（如总体均数、总体率）之差。如某地 100 名 40 岁以上正常男子血清胆固醇量的均数，一般不会恰好等于某地全体 40 岁以上正常男子血清胆固醇量的均数，两者之差就是抽样误差，产生抽样误差的基本原因，是总体内各观察单位存在着个体变异。由于总体内个体变异是必定存在的，因此，抽样误差是不可避免的。但是增加样本含量可以缩小抽样误差。

2. 随机测量误差　指的是由随机测量变异引起的误差。例如，分析天平的测得值就带有随机测量误差。由于观测中随机测量变异必定存在，因此，随机测量误差也是不可避免的。但是，改善测量手段和测量条件，可以将随机测量误差控制在更小的范围内。

偶然误差的性质：抽样误差与随机测量误差具有共同的性质，即其误差值一般较小；与真值相比其误差值可正可负，即方向是双向的；重复观测时其误差的出现，在一定范围内具有随机性，各次观察误差值的大小和方向相互独立。偶然误差是由多种对观察结果影响较小而又难以完全消除的因素综合影响的结果，它的出现是不可避免的，它表面上捉摸不定，实际上却有严格的统计规律性。

应该指出，偶然误差是有上述确定涵义的统计学概念，不要把它与偶然发生的误差相混淆。后者仅仅是指误差发生的方式是偶然的。实际上，有些偶然发生的误差却具有系统误差的性质。例如，由于操作的偶然失误，电压的突然变化，仪器的突然故障等所造成的误差等。

（二）系统误差

1. 产生系统误差的原因　有许多原因可引起系统误差，如使用未经校正的试验仪器或不符合规格的试剂；搜集资料的设计不周；观察方法、判断标准不统一或观察者主观偏见等。在相关学科中将这一类原因引起的误差统称之为偏倚（bias）。例如，使用未经校正的血压计测量血压；调查当地青年吸烟情况，以大中专学生为观察对象；将可疑病例作为确诊病例，将好转病例作为治愈病例进行统计分析等。

2. 系统误差的性质　虽然系统误差有各种各样，但它们都具有共同的性质。即其误差值与偶然误差值相比，一般较大；方向单一，或者偏大，或者偏小；在条件不变的情况下观察，同样大小和方向的误差将重复出现；具有某个（或几个）明确的原因（虽然

这样的原因有时尚未被发现)；当引起该系统误差的原因消除以后，该系统误差就不再出现；在观察中，系统误差会不会出现，出现的大小和方向等并无规律性。例如，使用未经校正的血压计测量血压量所造成的误差是较大的，方向是单一的，或夸大或缩小被测者的血压，只要血压计未校正，这样的误差一定会重复出现。血压计一经校正，这样的误差即消失，其原因是明确的，但当不了解血压计真实的情况时，人们将无法预言它是否会出现以及如果出现的话，它的大小和方向如何。偶然误差与系统误差性质的比较见表 1.1。

表 1.1　系统误差与偶然误差性质的比较

误差类型	大小	方向	大小和方向的重现性	产生的原因	可否避免	统计规律性
偶然误差	一般	双向	不一定	多种影响较小因素	不可避免	有
	较小		重现	综合影响的结果	但可控制	
系统误差	一般	单向	可重现	有少数确定	消除原因	无
	较大			的原因	即可避免	

3. 混杂因素引起的系统误差　如在分析不同职业与高血压患病率的关系时，即使各调查对象的血压测得值是正确的，但统计分析的结论仍可能偏离真实情况。这是由于不同职业人群的平均年龄可能差别很大(如运动员与工程师)，而平均年龄不同的人群高血压患病率差别是可以很大的。这时，年龄因素与职业因素对于血压的影响混在一起难以区分。年龄因素干扰了职业因素与高血压关系的分析。统计上把这种干扰因素称之为混杂因素，把因混杂因素干扰而造成统计结论偏离真实情况的现象称为混杂。统计资料中的混杂因素是客观存在的，并非由于观察者的失误而产生的。事物或现象间存在着广泛的、错综复杂的联系，当分析研究某些事物或现象间的联系时，应注意排除同时存在着的混杂因素的干扰。如何在统计设计和统计分析时控制和排除混杂因素的干扰，是卫生统计学的重要内容之一。有关这方面的内容见第五章、第八章、第十三章。

有一类误差是由于过失造成的，它也具有系统误差的性质，但它是一种非技术性的、责任性的错误，不包括在通常所说的系统误差之中。如读数错、计算错、记录错、录入错等。

八、参数和统计量

总体的指标称为参数，样本的指标称为统计量。例如，某地 40 岁以上的正常男子血清胆固醇量的总体均数，就是一个参数，而该地随机抽取的 100 名 40 岁以上的正常男子，其血清胆固醇量的均数，就是一个统计量。统计学约定参数用希腊字母表示，统计量用拉丁字母表示。如 μ 表示总体均数，π 表示总体率，σ 表示总体标准差，ρ 表示总体相关系数，$\bar{X}$ 表示样本均数，p 表示样本率，s 表示样本标准差，r 表示样本相关系数等。

九、统计推断

根据样本资料所提供的信息，对总体的特征作出推断，称为统计推断。统计推断包括两个方面：

1. 参数估计　根据样本资料所提供的信息，对总体指标的大小或所在范围作出估计为参数估计。又分为点估计和区间估计两种：①点估计是对总体指标作出一个定值的估计，虽然能给人一个明确的数量概念，但这只是一个近似值，常常不能满足实际工作的需要。②区间估计是估计总体参数所在的范围以及在这个范围内包含总体参数的可能性的大小。本书将在第三章和第五章叙述对总体均数和总体率的估计方法。

2. 假设检验　首先对总体指标作出一个假设，然后根据样本资料所提供的信息及有关统计量分布理论，对这个假设作出拒绝或不拒绝的判断。这种对于假设的拒绝或不拒绝的判断是具有概率性的，也就是说，这种判断的正确性不是百分之百的，即它是冒着犯有一定概率错误的风险作出判断的。然而这种判断比之于那种说不出判断错误概率的经验判断，要严密得多，可靠得多。因而假设检验作为一种经典的数据处理方法，早就成为自然科学和社会科学研究中的一种通用的方法。

假设检验有许多种，根据其所计算的统计量不同而命名，如 t 检验、U 检验、F 检验、χ^2（卡方）检验等。本书将以大量篇幅介绍各种假设检验方法。

复习思考题

1. 卫生统计学的任务是什么？为什么要学习卫生统计学？
2. 卫生统计工作分哪几个步骤，各步骤的意义与要求如何？
3. 观察值变异是由什么原因引起的？试举例说明。
4. 什么是观察单位？变量和变量值有何不同？
5. 统计资料分哪几种类型？区分统计资料类型的依据是什么？
6. 抽样研究的目的是什么？
7. 什么是统计推断，它包括哪两个方面？
8. 系统误差与偶然误差的性质如何？各举例说明。
9. 为什么说即使观察值是正确的，系统误差仍然可能发生？
10. 什么是混杂因素？它对统计结论有什么影响？

（福建卫生学校　周士楷）

第二章　数值资料的统计描述

第一节　数值资料的频数分布

频数就是观察值的个数。频数分布就是观察值在其所取值的范围内分布的情况。在观察值个数较多时，频数分布情况可用频数分布表和频数分布图来表示。编制频数分布表，绘制频数分布图，是整理资料的基本步骤之一。了解频数分布情况是研究数值资料的第一步。

一、频数分布表的编制

例 2.1　某市 150 名 3 岁女孩身高（cm）资料如下，试编制频数分布表。

80.1	100.1	97.0	96.7	97.9	100.7	86.2	91.7	94.7	90.8
82.5	102.6	99.1	96.6	99.3	85.2	89.2	90.6	95.1	93.6
84.4	104.8	101.3	98.7	101.5	87.1	89.0	92.7	96.8	92.7
87.2	83.5	103.2	101.6	84.4	88.4	91.8	93.6	99.2	94.4
89.3	84.2	82.3	84.5	87.9	89.4	91.9	94.5	86.9	95.6
89.1	86.5	85.0	87.6	89.3	90.4	92.1	95.0	89.3	96.3
91.3	89.7	87.4	89.8	88.7	90.2	92.9	97.2	91.4	90.3
90.5	88.9	88.1	88.2	91.1	93.0	95.6	98.7	90.0	93.5
92.4	90.0	88.0	90.7	91.7	93.8	94.4	87.3	93.9	92.8
92.6	90.0	90.8	90.1	93.2	94.4	97.3	89.0	92.9	94.3
94.7	92.8	90.3	92.8	93.6	94.8	98.3	88.5	94.0	96.0
94.8	92.3	93.3	93.1	95.1	97.0	84.5	91.1	94.3	93.4
97.1	95.8	93.7	95.1	94.9	99.4	86.4	91.7	96.5	92.5
96.2	94.3	94.2	94.6	96.4	100.9	89.1	93.2	98.4	95.5
99.5	97.5	95.1	96.2	99.5	85.7	88.4	92.5	91.1	97.3

编制频数分布表步骤如下：

1. 计算全距　找出观察值中最大值与最小值，两者之差即为全距。本例最大值为 104.8，最小值为 80.1，全距＝104.8－80.1＝24.7（cm）。

2. 确定组段数、组距和组段　根据全距的大小和观察值个数多少，决定组段数。全距大，观察值个数多，组段数可适当多些。一般取 10～15 个组段为宜。组段数过多，编制过程和计算较繁，组段数过少，计算误差较大。本例组段数暂定为 12。

根据组段数和全距，决定组距。组距＝全距/组段数。本例组距为 24.7/12＝2.06，为归组和计算方便，取组距为 2cm。

划分组段是将观察值依次划分为若干个段落，这些段落称为组段。各组段的起点为下限，终点为上限。第一组段要包括最小的观察值，即该组段的下限应略小于或等于最小观察值；最后一组段要包括最大观察值，即该组段的上限应略大于或等于最大观察值。

3. 列表归组　列出频数分布表（表 2.1）。表中第（1）栏为组段，连续性资料各组段的上限不标出，以表示资料的连续性，并有利于归组汇总。用划记法或分卡法将各观察单位分别归入各组段，得第（2）栏。清点各组段内的观察值个数即得各组段频数，将各组段频数填入第（3）栏，合计各组段频数为总频数。

表 2.1　某市 150 名 3 岁女孩身高频数分布

组段（cm） （1）	划　记 （2）	频数 （3）
80～	一	1
82～	下	3
84～	正下	8
86～	正正	10
88～	正正正正	19
90～	正正正正下	23
92～	正正正正正	26
94～	正正正正正	24
96～	正正正下	17
98～	正正	10
100～	正一	6
102～	下	2
104～106	一	1
合　计		150

在计算机普遍应用的今天，频数表的编制一般由计算机进行。用计算机编制频数表准确、快速，可根据用户的需要，随时变换组距，以输出理想的频数表。但即便使用计算机，也应保证原始数据的准确输入和组距的设计合理。

二、频数分布图

为了更加直观地了解频数分布情况，通常在编制频数表的基础上，绘制频数分布图。常见的频数分布图有两种，即直方图和多边图。和编制频数表一样，频数分布图亦可由计算机绘制。手工绘制频数分布图的方法见第十二章。

图 2.1 和图 2.2 是根据表 2.1 资料绘制的直方图和多边图。

三、频数分布类型

从频数分布的图形来看，常见频数分布有三种类型：

1. 正态分布型　如图 2.3（1），整个图形以高峰所在处的垂线为中心，左右两侧逐渐下降并对称。在这类分布中，以正态分布为最典型（详后），例 2.1 资料即属此类型分布。

2. 正偏态分布型　如图 2.3（2），整个图形不对称，高峰偏左，即观察值较小的这一端，集中了较多的频数。属于这类分布的资料并不少见，如正常人体中某些非必需元素含量的频数分布；一些传染病潜伏期的频数分布；粉尘粒子大小的频数分布等。

3. 负偏态分布型　如图 2.3（3），整个图形不对称，高峰偏右，即观察值较大的这

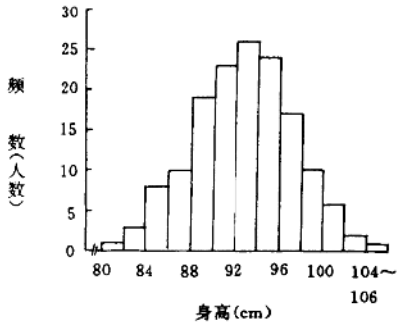

图 2.1　某市 150 名 3 岁女孩身高频数分布直方图

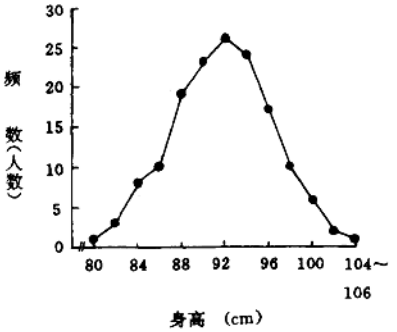

图 2.2　某市 150 名 3 岁女孩身高频数分布多边图

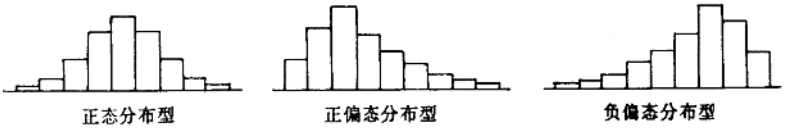

图 2.3　几种不同类型的频数分布示意图

一端，集中了较多的频数。属于这类分布的资料较为少见，如某班学生成绩的频数分布，多数学生得分较高，少数学生得分较低；又如儿童少年视力频数分布，多数学生视力较好，少数学生视力较差。

第二节　集中趋势指标

从频数表和频数分布图中，可以大致看出频数分布的规律，也可以大致了解频数分布的特征——集中趋势和离散程度，但这种认识是粗略的。如欲准确掌握频数分布的特征，就应作频数分布特征的定量描述，即计算集中趋势和离散程度指标。

集中趋势指标又称平均数，它反映了观察值的集中位置或平均水平。也可以说，是观察值的典型水平或代表值。常用的集中趋势指标有算术均数（均数）、几何均数和中位数等。

一、算术均数（均数）

（一）应用条件

均数最适用于对称分布，尤其是正态分布资料。因为这时均数位于中央，能反映观察值的集中趋势。当观察值个数较少，而其频数分布基本对称或从专业上可推断分布为正态或近似正态者，也可用均数作为其集中趋势指标。

（二）计算方法

1. 直接法　按式（2.1）计算。

$$\bar{X}=\frac{\Sigma X}{n}=\frac{X_1+X_2+X_3+\cdots+X_n}{n} \qquad (2.1)$$

式中 $\bar{X}$ 为均数，（读作 eksba），X_1，X_2，X_3，$\cdots X_n$ 为各观察值，$\sum$ 为求和符号，