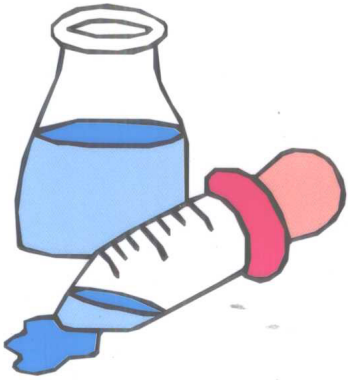

实用卫生统计

主编 许国章

復旦大學出版社

实用卫生统计

主　编　许国章

副主编　董红军　张　涛

编　委（按章节先后排序）

许国章　边国林　马　瑞　周爱明

復旦大學出版社

图书在版编目(CIP)数据

实用卫生统计/许国章主编. —上海:复旦大学出版社,2010.6
ISBN 978-7-309-07181-8

Ⅰ. 实… Ⅱ. 许… Ⅲ. 卫生统计 Ⅳ. R195.1

中国版本图书馆 CIP 数据核字(2010)第 058769 号

实用卫生统计
许国章 主编
出品人/贺圣遂 责任编辑/贺 琦

复旦大学出版社有限公司出版发行
上海市国权路 579 号 邮编:200433
网址:fupnet@fudanpress.com http://www.fudanpress.com
门市零售:86-21-65642857 团体订购:86-21-65118853
外埠邮购:86-21-65109143
上海申松立信印刷有限责任公司

开本 850×1168 1/32 印张 9.875 字数 227 千
2010 年 6 月第 1 版第 1 次印刷

ISBN 978-7-309-07181-8/R·1139
定价:21.00 元

前言 FOREWORD

卫生统计学是运用概率论与数理统计的原理及方法，结合卫生实际，研究数字资料的收集、整理分析与推断的一门学科，是每一位流行病学工作者必须具备的重要工具和研究手段。流行病学工作者在暴发疫情中识别危险因素、评价卫生措施效果、进行科学决策，均涉及大量的统计学知识，需要灵活应用统计学方法，解决实际问题。但现有的卫生统计学类教材往往面向高校学生，理论性强，基层流行病学工作者阅读起来有一定困难，因此编写一本通俗易懂、实用性强、操作性强、紧密联系基层疾病控制工作实际的统计学类教材显得十分必要。鉴于此，我们组织多年从事疫情分析、疫情暴发调查、慢性病统计的专家编写了《实用卫生统计》一书，力求深入浅出地阐释统计学原理，强调理论与实际相结合，介绍适用于基层对疾病暴发调查、统计分析的统计软件使用方法。

本书共分12章，其中第一至二章、第六至十一章深入浅出地介绍卫生统计学基础理论和常用的统计方法，包括统计描述和概率分布、总体均数的估计和假设检验、χ^2检验、方差分析、秩和检验、简单线性回归分析、寿命表和生存分析等。第三至第四章着重介绍现场调查设计、数据录入和整理，包含在现场调

查中经常涉及的样本含量的估计及 Epi Data 的使用等。第五章《统计图表的制作》结合疾控系统疫情分析人员每天使用的“中国疾病预防控制信息系统”，介绍如何应用 Excel 对国家网络直报数据库的数据进行直观的描述和充分的分析利用。第十二章《Epi Info 2002 软件应用》，运用大量的实例介绍如何灵活应用 Epi Info 软件，使得现场调查、暴发疫情调查中的统计分析更加快捷准确。本书遵循实用性、科学性、指导性并重的编写原则，主要突出实际应用，书中所举事例大多取自实地疾病监测的数据，适用于各级疾病预防控制机构和基层预防保健工作人员学习和参考。

由于知识水平有限，错漏、不足之处在所难免，恳请各位读者不吝提出宝贵意见。

编　者

2010 年 6 月

目录 CONTENTS

第一章 绪论 …… 1

第一节 卫生统计学简介 …… 1

第二节 统计工作的步骤 …… 3

第三节 统计资料的类型 …… 5

第四节 统计学中的几个基本概念 …… 6

第五节 使用卫生统计学应注意的问题 …… 12

第二章 统计描述和概率分布 …… 13

第一节 计量资料的统计描述 …… 13

第二节 计数资料的统计描述 …… 21

第三节 率的标准化法的意义和计算方法 …… 27

第四节 常用概率分布 …… 33

第三章 现场调查设计 …… 41

第一节 调查设计的意义 …… 41

第二节 调查设计的基本原则与步骤 …… 42

第三节 抽样设计 …… 50

第四节 样本含量的估计 …… 65

第五节　现场调查设计中应考虑的问题…… 68

第四章　数据录入和整理…… 79
第一节　数据的审核…… 79
第二节　数据的编码及录入…… 84
第三节　数据整理…… 88
第四节　Epi Data 的使用…… 94

第五章　统计图表的制作…… 113
第一节　统计表…… 113
第二节　统计图…… 116
第三节　用 Excel 制作统计图表…… 128

第六章　总体均数的估计和假设检验…… 148
第一节　抽样研究与抽样误差…… 148
第二节　t 分布…… 152
第三节　总体均数的估计…… 154
第四节　假设检验的基本步骤…… 157
第五节　样本与总体比较的假设检验…… 159
第六节　配对设计资料的假设检验…… 160
第七节　两样本比较的假设检验…… 162
第八节　第 1 类错误与第 2 类错误…… 163

第七章　χ^2 检验…… 165
第一节　四格表资料的 χ^2 检验…… 165
第二节　行×列表(R×C)的 χ^2 检验…… 171
第三节　配对设计的 χ^2 检验…… 172
第四节　其他行×列(R×C)表资料的统计分析…… 174

第八章　方差分析 …… 179
第一节　方差分析的基本思想 …… 180
第二节　方差分析的适用条件 …… 185
第三节　不同设计资料的方差分析 …… 192

第九章　秩和检验 …… 205
第一节　配对设计资料的符号秩和检验
（Wilcoxon 符号秩和检验法） …… 206
第二节　完全随机设计两组独立样本的秩和检验 …… 210
第三节　完全随机设计多组独立样本的秩和检验 …… 216
第四节　随机化区组设计资料的秩和检验 …… 218
第五节　多个样本间两两比较的秩和检验 …… 220

第十章　简单线性回归分析 …… 225
第一节　简单线性回归的概念 …… 225
第二节　简单线性回归方程的求法 …… 229
第三节　线性回归方程的检验 …… 232

第十一章　寿命表和生存分析 …… 238
第一节　寿命表 …… 238
第二节　生存分析 …… 254

第十二章　Epi Info 2002 软件应用 …… 274
第一节　Epi Info 软件简介 …… 274
第二节　Epi Info 软件数据分析中的基本操作 …… 280
第三节　Epi Info 软件数据的常见统计分析 …… 288
第四节　常用流行病学计算 …… 304

第一章
绪　论

第一节　卫生统计学简介

卫生统计学是运用概率论与数理统计的原理及方法，结合卫生实际，研究数字资料的收集、整理分析与推断的一门学科。

卫生研究的对象主要是人体以及与人体健康有关的各种因素。生物现象的一个重要特点就是普遍存在着变异。所谓变异（个体差异），系指相同条件下同类个体之间某一方面发展的不平衡性，系偶然因素起作用的结果。例如同地区、同性别、同年龄的健康人，他们的身高、体重、血压、脉搏、体温、红细胞数、白细胞数等数值都会有所不同。又如在同样条件下，用同一种药物来治疗某病，有的患者被治愈，有的疗效不显著，有的可能无效甚至致患者死亡。引起客观现象差异的原因是多种多样的。归纳起来，一类是普遍的、共同起作用的主要因素；另一类则是偶然的、随机起作用的次要因素。这两类原因总是错综复杂地交织在一起，并以某种偶然性的形式表现出来。科学的任务就在于，要从看起来是错综复杂的偶然性中揭露出潜在的必然性，即事物的客观规律性。这种客观规律性是在大量现象中发现的。比如临床要观察某种疗法对某病的疗效时，如果观察的患者很少，便不易正确判断该疗法对某病是否有效；但当观察患者

的数量足够多时，就可以得出该疗法在一定程度上有效或无效的结论。所以，卫生统计学是卫生科学研究的重要工具。

卫生统计学在20世纪20年代以后才逐渐形成为一门学科。解放前，我国学者即致力于把统计方法应用到卫生工作中去，但人力有限、范围较窄。解放后，随着卫生科研工作的发展，卫生统计方法得到迅速普及与提高。通过大量实践，在不少方面积累了自己的经验，丰富了卫生统计学的内容。而电子计算机的使用，更促进了多变量分析等统计方法在卫生研究中的应用。

卫生统计学的内容包括：①统计研究设计。我们制订调查计划或实验设计时，除专业问题外，还必须从卫生统计学的角度考虑，使调查或实验结果能够科学地回答所研究的问题。一个好的设计可以用较少的人力、物力和时间取得更多的较可靠的资料。②总体指标的估计。卫生研究中实际观测或调查的部分个体称为样本，研究对象的全体称为总体。人们除用均数、率等统计指标对调查或实验结果进行描述外，更重要的是通过样本的信息，估计总体中相应的统计指标，即参数估计。③假设检验。就是依据资料性质和所需解决的问题，先建立适当的假设，然后采用适当的检验方法，根据样本是否支持所作的假设，决定对假设的接受或拒绝。④联系、分类、鉴别与鉴测等研究。在疾病的防治工作中，经常要探讨各种现象数量间的联系，寻找与某病关系最密切的因素；要进行多种检查结果的综合评定，探讨疾病的分型分类，选择治疗方案；要对某些疾病进行预测预报、流行病学监督；对药品制造、临床检验工作等作质量控制，以及卫生人口学研究等。卫生统计学，特别是其中的多变量分析，为解决这些问题提供了必要的方法和手段。

作为卫生科学工作者，学习和掌握一定的统计学知识是十分必要的。第一，在阅读卫生学书刊中，经常会遇到一些统计学方面的名词概念，有了这方面的知识，有助于正确理解文章的涵

义；第二，在实际工作中，经常要做登记工作，要填写各种报表，只有懂得原始登记与统计结果的密切关系，并掌握收集、整理与分析资料的基本知识与技能，才能自觉、认真地把登记工作做好，积累有科学价值的资料；第三，参加科研工作时，从开始设计到数据整理分析与统计结果的表达，每一步骤都需要统计学知识；第四，在制订计划、检查工作、总结经验时，都离不开统计数字，尤其在撰写科研论文时，有了统计学知识，才能使数据与观点密切结合，作出正确的结论。

医务工作者学习统计学，首先必须明确：我们应该掌握的关键不是数学原理，而是怎样合理、恰当地把数理统计的方法应用到卫生科研工作中去，并结合专业知识，提高分析问题与解决问题的能力。其次在学习过程中，要理论联系实际，重视实习与练习。作业中要遵守数学上的规则与习惯，如小数点及各个位数应上下对齐，一个多位数的数值不能分写成两行，等号不能写在一行的末了而应写在第二行的开头等等。再次，各种统计符号必须写正确，汉字、阿拉伯字与外文字母必须写清楚，只有在学习时养成良好的习惯，将来工作中才能少出差错。

最后我们着重指出：统计工作最根本的一条就是实事求是，如实反映情况。因此，在日常工作或科学研究中，必须养成严肃认真的作风和反复核对的习惯，同一切弄虚作假的现象进行坚决的斗争，尽最大努力获得正确数据，使分析结论建立在可靠的基础上。

第二节 统计工作的步骤

统计学对统计工作的全过程起指导作用，任何统计工作和统计研究的全过程都可分为以下 4 个步骤。

1. 设计　在进行统计工作和研究工作之前必须有一个周密的设计。设计是在广泛查阅文献、全面了解现状、充分征询意见的基础上，对将要进行的研究工作所做的全面设想。其内容包括：明确研究目的和研究假说，确定观察对象、观察单位、样本含量和抽样方法，拟定研究方案、预期分析指标、误差控制措施、进度与费用等。设计是整个研究工作中最关键的一环，也是指导以后工作的依据。

2. 收集资料　遵循统计学原理采取必要措施得到准确可靠的原始资料，及时、准确、完整是收集统计资料的基本原则。卫生工作中的统计资料主要来自以下 3 个方面。①统计报表：是由国家统一设计，有关医疗卫生机构定期逐级上报，提供居民健康状况和医疗卫生机构工作的主要数据，是制订卫生工作计划与措施、检查与总结工作的依据。如《法定传染病报表》、《职业病报表》、《医院工作报表》等。②经常性工作记录：如卫生监测记录、健康检查记录等。③专题调查或实验。

3. 整理资料　收集来的资料在整理之前称为原始资料，原始资料通常是一堆杂乱无章的数据。整理资料的目的就是通过科学的分组和归纳，使原始资料系统化、条理化，便于进一步计算统计指标和分析。其过程是：首先对原始资料进行准确性审查（逻辑审查与技术审查）和完整性审查；再拟定整理表，按照“同质者合并，非同质者分开”的原则对资料进行质量分组，并在同质基础上根据数值大小进行数量分组；最后汇总归纳。

4. 分析资料　其目的是计算有关指标，反映数据的综合特征，阐明事物的内在联系和规律。统计分析包括统计描述和统计推断。前者是用统计指标与统计图（表）等方法对样本资料的数量特征及其分布规律进行描述；后者是指如何抽样，以及如何用样本信息推断总体特征。进行资料分析时，需根据研究目的、设计类型和资料类型选择恰当的描述性指标和统计推断方法。

统计工作的4个步骤紧密相连、不可分割，任何一步的缺陷，都将影响整个研究结果。

第三节 统计资料的类型

卫生统计资料按其性质一般分为计数资料与计量资料两类。不同类型的统计资料应采用不同的统计分析方法。

计数资料是先将观察单位按某种属性或类别分成若干组，再清点各组，观察单位个数所得到的资料。如临床某些检验结果用阳性或阴性反应表示，对一批某病患者检验完毕后，清点呈阳性或阴性反应的各有若干例。又如要调查某人群的血型分布，先按A、B、AB、O 4型分组，再清点各血型组人数。计数资料每个观察单位之间没有量的差别，但各组之间具有质的不同，不同性质的观察单位不能归入一组。对这类资料通常是先计算百分比或率等相对数，需要时做百分比或率之间的比较，也可做两事物之间相关的相关分析。

计量资料是用仪器、工具或其他定量方法对每个观察单位的某项标志进行测量，并将测量结果用数值大小表示出来的资料，一般带有度量衡或其他单位。如检查一批应征青年体重，需要磅秤测量，通常以千克(kg)为单位，测得许多大小不一的体重值。其他如身高(cm)、血压(mmHg)、脉搏(次/分)、红细胞数($\times 10^{12}$/L)、转氨酶(u)等，都属于计量资料。每个观察单位的观测值之间有量的区别，但同一批观察单位必须是同质的。对这类资料通常先计算平均数与标准差等指标，需要时做各均数之间的比较或各变量之间的分析。

还有一些资料，也是将观察单位按某种属性或某个标志分组，然后清点各组观察单位个数得来的，但所分各组之间具有等

级顺序。这些资料既具有计数资料的特点，又兼有半定量的性质，称为等级资料或半定量资料。例如某病住院患者的治疗结果，按治愈、好转、无效、死亡分组，同样各组之间具有顺序与程度之别。分析等级资料常用的统计指标有比和率，常用的统计方法有秩和检验、参照单位分析等。

在卫生工作实践中，根据分析研究的目的，计数资料与计量资料可以互相转化。例如血压值本是计量资料，但如果将一组20～40岁成年人的血压值分为血压正常与血压异常两组，再清点各组人数，于是这组血压计量资料（血压值）就转化成为计数资料了。假若将这组血压计量资料（血压值）按低血压（＜80/60 mmHg）、正常血压（80～130/60～89 mmHg）、轻中度高血压（＞130/90～110 mmHg）、重度高血压（＞130/＞110 mmHg）的等级顺序分组，清点各组人数，这时这组计量资料（血压值）又转化为等级资料了。

由于计量资料可以得到较多的信息，所以凡能计量的，尽量采用计量资料。

第四节　统计学中的几个基本概念

一、样本与总体

前面已提及，卫生研究中实际观测或调查的一部分个体称为样本，研究对象的全部称为总体。如作水质检验时从井水或河水中采的水样，临床检验中从患者采取的血液或其他活体组织标本，是样本；而整个一口井或一条河的某一段所有的水，某患者全身所有的血液或某个组织器官，则是总体。这类总体是具体存在的，但另有些总体却是假想的，只是理论上存在的一个

范围。例如试验某一治疗流感新药的疗效，最初接受治疗的一批流感患者，不论数量多少，都只是一个样本。若该药疗效得到肯定，从而加以推广，那么此后凡在相同条件下接受该药治疗的所有流感患者，都属于这个总体。可是当初试用时，这个总体还不存在，是假想的。

总体包含的观察单位通常是大量的甚至是无限的，在实际工作中，一般不可能或不必要对总体中的每个观察单位逐一进行研究。我们只能从中抽取一部分观察单位加以实际观察或调查研究，根据对这一部分观察单位的观察研究结果，再去推论和估计总体情况。如上述某新药治疗流感，试验治疗的只是少数有限的患者，而结论却要推广到全体患者，得出一个该药对所有流感患者疗效的规律性认识。所以说，观察样本的目的在于推论总体，这就是样本与总体的辩证关系。

为了使样本能够正确反映总体情况，对总体要有明确的规定，总体内所有观察单位必须是同质的；在抽取样本的过程中，必须遵守随机化原则；样本的观察单位还要有足够的数量。

二、概率

概率又称机率，是用以描述某事件发生的可能性大小的一个数值。

在自然界和人类社会中，存在着两类不同的现象：①在一定条件下，肯定发生的事件叫做必然事件，肯定不发生的事件叫做不可能事件。如在适当温度、湿度下经一定时间孵化，正常受精鸡蛋必然会孵出小鸡来，而石头是不可能孵出小鸡来的。必然事件与不可能事件虽然形式相反，但两者在发生某种结果与否都是确定的，故统称确定性现象。②在基本条件不变的情况下，可能发生的结果有多种，究竟发生哪种结果，事先不能肯定，这类现象叫做随机现象。随机现象的表现结果称为随机事件。

如任意抛掷一枚硬币，可能徽花向上也可能币值向上，抛掷前不能肯定，这是一个随机现象，而结果出现“徽花向上”则是一个随机事件。

1. *古典概率* 是最简单的随机现象的概率计算。这类随机现象具有两个特征：① 在观察或试验中它的全部可能结果只有有限个，譬如为 n 个，记为 E_1，E_2，…，En，而且这些事件是两两互不相容的，即任何两个事件不能同时发生；② 事件 E_1，E_2，…，En 的发生或出现是等可能的，即它们发生的概率都一样。古典概率的大部分问题都能形象地用摸球模型来描述，有利于直观地理解概率论的许多基本概念；而且它有着多方面的重要应用，例如工业产品的抽样检查等。

2. *统计概率* 上述“事件”是指不能再进行分解或不能由其他事件构成的基本事件。在实际工作中，基本事件的发生并不总是等可能的，而且有时为无穷多个。这样就有必要把古典概率的定义加以推广，从事后经验的角度来理解概率的意义。实践证明，虽然个别随机事件在某次试验或观察中可以出现也可以不出现，但在大量重复试验中它却呈现出明显的规律性。假设在相同条件下，独立地重复做 n 次试验，某随机事件 A 在 n 次试验中出现了 m 次，则比值 m/n 称为随机事件 A 在 n 次试验中出现的频率。当试验重复很多次时，随机事件 A 的频率 m/n 就会在某个固定的常数 P 附近摆动，而且 n 愈大摆动的幅度愈小。即试验次数(n)越多，A 的频率(m/n)变动越小，越趋稳定。这种规律性称为统计规律性。频率的稳定性说明随机事件发生的可能性大小是随机事件本身固有的、不随人们意志为转移的客观属性。所以在卫生科研中，当 n 充分大时，就以频率作为概率的近似值，记作 $P(A)$。

由此可见，频率是就样本而言的，而概率是从总体的意义上而言的。这样，概率就为预计某一事件发生的可能性大小提供

了衡量的尺度。例如：某病患者 40 例，用某疗法治疗后，其中 35 例痊愈，治愈者占治疗人数的 35/40(87.5%)，这就是频率。因为数量少，这个频率可能波动较大。假如经过长期的大量观察，比如数百、数千例，得到治愈率为 70%，我们就可以说，该疗法治愈某病的概率近似值为 70%。又如：某院妇产科在 1 个月内出生婴儿 30 名，其中男婴 18 名，占新生儿数的 18/30(60%)，称为频率。大量统计表明，人口中男女的比例基本上是 1∶1。这是个较稳定的常数，即概率的近似值。于是，在婴儿分娩前，我们就可用它作为尺度，预计是男的概率为 1/2(0.5 或 50%)，是女的概率也为 1/2(0.5 或 50%)。

通过以上讨论，可以知道：如果某事件是必然事件，则有 $m = n$，所以必然事件的概率等于 1；如果某事件是不可能事件，则有 $m = 0$，所以不可能事件的概率等于 0；如果某事件是随机事件，则有 $0 < m < n$，所以随机事件的概率是介于 0 与 1 之间的一个数。某事件的概率越接近 0，表示发生的可能性越小；越接近 1，表示发生的可能性越大。

三、随机变量

简单地说，是指随机事件的数量表现。例如一批注入某种毒物的动物，在一定时间内死亡的只数；某地若干名男性健康成人中，每人血红蛋白量的测定值；等等。另有一些现象并不直接表现为数量，例如人口的男女性别、试验结果的阳性或阴性等，但我们可以规定男性为 1，女性为 0，则非数量标志也可以用数量来表示。这些例子中所提到的量，尽管它们的具体内容是各式各样的，但从数学观点来看，它们表现了同一种情况，这就是每个变量都可以随机地取得不同的数值，而在进行试验或测量之前，我们要预言这个变量将取得某个确定的数值是不可能的。

按照随机变量可能取得的值，可以把它们分为两种基本类