全国高等医学院校教材

卫生统计学

（第2版）

主编 王 燕
安 琳

北京大学医学出版社

卫生统计学

[redacted]

[redacted]

全国高等医学院校教材

卫生统计学

（第2版）

主　　编　王燕　安琳

副 主 编　罗树生

编写人员（按顺序笔画为序）

王晓莉　王　燕　任正洪

安　琳　易伟宁　罗树生

高燕秋　康楚云

北京大学医学出版社

图书在版编目（CIP）数据

卫生统计学（第2版）/王燕 安琳主编．—北京：北京大学医学出版社，2008
ISBN 978-7-81116-312-4

Ⅰ．卫… Ⅱ．①王…②安… 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（2008）第154801号

卫生统计学（第2版）

主　　编：王　燕　安　琳
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
印　　刷：北京东方圣雅印刷有限公司
经　　销：新华书店
责任编辑：暴海燕　**责任校对**：杜悦　**责任印制**：郭桂兰
开　　本：787mm×1092mm　1/16　**印张**：11　**字数**：267千字
版　　次：2009年1月第2版　2009年1月第1次印刷　**印数**：1—5000册
书　　号：ISBN 978-7-81116-312-4
定　　价：16.90元

第2版前言

卫生统计学是一门关于如何运用数理统计学的原理和方法进行医学科研，特别是关于如何收集、整理、分析和解释医学数据的应用科学。本书的编者具有多年教授卫生统计学的教学经验，了解初学者，特别是成人学生学习卫生统计学的重点和难点，在编写时我们采用深入浅出的讲解，充分发掘和调动学生日常生活和工作经验去理解统计中的各种概念和原理，而去掉了传统卫生统计学大量的抽象概念和数理统计公式的推导。本书编写力求通俗易懂、简洁实用，因此，对于已具有一定医疗卫生工作经验的人员学习卫生统计学的基本知识以及解决科研中的实际问题具有一定的参考价值。

随着计算机的普及，为了使读者从繁重的手工计算中解脱出来，本书增加了利用SPSS统计软件来完成统计分析的计算过程。SPSS软件是目前国际上流行的统计分析软件，为了帮助读者掌握该软件的使用方法，本书特意编写了一章SPSS软件使用方法介绍，而且在其他各章中只要涉及可以用SPSS软件实现的统计计算，均在相应的章节中给出如何用SPSS统计软件完成这些统计分析的指导，同时也给出了SPSS输出结果的形式和对结果的解释，极大地方便了读者的学习和应用。

限于编者的学识和精力，本书难免有不足之处，我们将虚心吸取广大读者和专家的批评与建议，争取在再版中给予弥补。

编者：安琳　王燕

2008年6月于北京

目　录

第一章　绪论

第一节　卫生统计学简介

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的应用，是关于医学研究中资料的收集、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。为什么要学习卫生统计学？简言之是为了进行医学领域的科学研究和科学决策。

卫生统计学的基本内容包括三个方面：①卫生统计学的基本原理和方法，包括医学科研设计和数据处理的基本统计理论和方法。②健康统计，包括医学人口统计、疾病统计、生长发育统计等。③卫生服务统计，包括卫生资源、卫生服务等统计。根据教学目的的需要，本教材主要包括：卫生统计学的基本原理和方法中的数据处理的基本统计理论和方法；健康统计中的医学人口统计和疾病统计。

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、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），亦称观察值或指标值。例如，为了了解某地区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“阳性”或“阴性”。

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岁男孩的身体发育状况，这时，一个2岁男孩是观察单位，测量指标为身高（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、u 检验、方差分析、直线相关与回归等（见第五、六、七章）。

（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，研究得到的每个儿童卡疤“阳性”或“阴性”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“阳性”组和“阴性”组，然后清点每组中的人数，对这样得到的数据做进一步分析。描述分类变量常用的统计指标有率、构成比等（见第三章），统计分析方法有 u 检验、χ^2 检验（见第八章）。

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：文盲、小学、初中、高中和大专及以上等；如疗效，可按疗效的不同程度，分为治愈、显效、有效和无效。

（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；≥2500克定义为非低出生体重两类，这时就成了二项分类的变量。如果分组再细一些，将出生体重为＜2500克定义为低出生体重，2500至＜4000克定义为正常出生体重，≥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，文化程度是按文盲、小学、初中、高中、大专及以上分组，可分别赋值为0、1、2、3、4。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，同质是指是同一地区、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原由。

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项变量值的集合。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，欲了解某地区2005年4月20日时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。随机抽样的概念和具体方法见下段。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，它的总体就是所有可能影响疗效的情况与样本相同的胃溃疡病人。

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特征，正是卫生统计学的奥妙所在。

五、随机抽样（random sampling）

随机抽样是指总体中每个观察单位都有一个不为零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。常用的随机抽样方法有下面四种。

（一）单纯随机抽样（simple random sampling）

即先将调查总体的全部观察单位编号，再用抽签或产生随机数字等方法随机抽取部分观察单位作为样本。单纯随机抽样是最基本的抽样方法，是其他抽样方法的基础。该方法要求对被抽样的总体中的每个观察单位编号，需要观察单位的名册。

（二）系统抽样（systematic sampling）

又称间隔抽样或机械抽样。将总体的所有观察单位按一定顺序排列，按照一定间隔抽取一个观察单位，进而抽取若干个观察单位组成样本。例如，某班有100个学生，需要10个学生作为样本，抽样间隔为10。首先，将100个学生按照学号排列，然后，在1、2、……9、10数字中用单纯随机抽样抽取一个数字，假定抽中的是数字3，则排在第3位、第13位、……、第83位、第93位的同学被抽中作为样本。该方法要求对被抽样的总体中的所有

观察单位排序。

（三）分层抽样（stratified sampling）

即先按某种特征将总体分为组别或类别，统计上叫“层”，再从每一层内进行随机抽样。例如，先将100个学生按照性别分为“男”、“女”两层，再按上述单纯随机抽样或系统抽样方法抽取样本。

（四）整群抽样（cluster sampling）

调查总体中本身存在小“群体”，用随机抽样的方法抽取若干个“群体”，这些被抽中的“群体”包括的所有观察单位作为样本。例如，100个学生中有10个小组，将“小组”作为“群体”，给10个小组编号为1、2…9、10，在1～10数字中用单纯随机抽样抽取一个数字，该数字代表的小组中的所有学生作为样本。

六、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、随机测量误差及抽样误差，抽样误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统计指标间仍会存在差别，称这种差别为抽样误差。抽样误差是由于个体差异及抽样造成的，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（μ）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（μ）间有一定差别，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。

七、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 $0 \leqslant P \leqslant 1$，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。如用某药治疗 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，若经过多次试验或对许多人的治疗，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 $P \leqslant 0.05$ 称为小概率事件，表示某事件发生的可能性很小，是几乎不可能发生的事件。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，具体包括研究的意义、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、整理资料方法、分析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，若确实不详可填写“不详”。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？具体的样本量计算方法未被包含在这本书中，可参见其他卫生统计学教材。

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。

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每一份原始数据，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，性别变量取值范围为“1”、“2”或“男”、“女”；出生体重变量取值范围为1500g－6000g，等等；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“14岁”这样结婚年龄很低的，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“25年”，这显然是不可能。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。

本书中主要介绍分析资料的方法，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的基础上，介绍了 SPSS 软件的操作。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。

习　题

1. 卫生统计学与医学的关系是什么？
2. 变量的类型有哪些？各有什么特点？有什么联系？分清变量类型有什么意义？
3. 请说出同质与变异的基本概念，并结合例子给以解释。
4. 请举例说明总体和样本的基本概念。抽样研究的目的是什么？思考：某医生用某药治疗了前来就医的 200 例胃溃疡患者，这 200 例胃溃疡患者的治疗效果是不是样本？是否是随机抽样？它的总体是什么？研究的目的是什么？
5. 误差的概念及分类是什么？抽样误差产生的原因是什么？
6. 卫生统计工作的基本步骤是什么？
7. 收集资料时，对资料的要求是什么？

（王燕）