YIXUE TONGJIXUE

医学统计学

主　编　王洁贞　李颖琰　陈冠民

郑州大学出版社

医学统计学

主　编　王洁贞　李颖琰　陈冠民

郑州大学出版社

图书在版编目(CIP)数据

医学统计学/王洁贞,李颖琰,陈冠民主编. —郑州:郑州大学出版社,2002.1
ISBN 7-81048-537-7

Ⅰ.医… Ⅱ.①王…②李…③陈… Ⅲ.医学统计 Ⅳ.R195.1

中国版本图书馆 CIP 数据核字(2001)第 095060 号

郑州大学出版社出版发行
郑州市大学路 40 号　　邮政编码:450052
出版人:谷振清　　发行部电话:0371-6966070
全国新华书店经销
郑州市毛庄印刷厂印制
开本:787 mm×1 092 mm　　1/16
印张:20
字数:474 千字
版次:2002 年 2 月第 1 版　　印次:2003 年 4 月第 2 次印刷

书号:ISBN 7-81048-537-7/R·486　　定价:34.00 元

编写人员名单

主　编　王洁贞　李颖琰　陈冠民

副主编　施学忠　薛付忠　陈东峨　程光文

编　委　（按姓氏笔画排列）

丁守銮　王洁贞　王爱英　艾　东

迟玉聚　张合喜　刘言训　李颖琰

陈冠民　施学忠　韩　兢　韩定芬

程光文　薛付忠

编写说明

为了提高医学专业硕士研究生独立进行科学研究的能力，加强研究生课程建设，应广大医学专业硕士研究生的要求，由山东大学、郑州大学、武汉大学、武汉科技大学、新乡医学院、皖南医学院几所院校联合编写了这本《医学统计学》教材。

本书在保证“三基”内容的基础上，注重加强统计思维方法、统计分析和独立进行科学研究的能力培养。例如，加强了实验设计及其分析方法（第8～11章）和SAS软件在常用多元分析中应用（第15～18章）的内容。全书共18章：第1章绪论，第2～14章为基本统计方法，第15～18章为医学常用多元分析方法。

在本书编写过程中，山东大学公共卫生学院卫生统计学教研室的吴学森博士，贾红英、刘云霞、由智勇硕士，郑州大学公共卫生学院卫生统计学教研室的韩耀风、周莹硕士，为本书稿在微机上的修改、打印和编排做了大量烦琐、细致的工作。在此谨致以衷心的感谢。

编写过程中，全体编者努力工作，力图提高教材质量。但限于编者的水平，书中难免还有缺点和错误。欢迎使用本书的师生和读者批评指正。

王洁贞　李颖琰　陈冠民

2001.11.6

目　录

第一章　绪　论

第一节　医学统计学的作用和内容

世界上各类现象的发展变化规律，都表现为质与量的辩证统一。要认识某现象客观存在的规律性，就必须认识其质与量的辩证关系，认识其数量关系的特征及度的界限，这一切都离不开统计学。

统计学（statistics）是认识社会和自然界中随机现象的数量特征的一门科学。

自然界存在的各种现象可归纳为两类现象。一类是在一定条件下必然发生的现象，称为必然现象。这类现象的发生在一定条件下是确定性的。例如在地球上，上抛的石子必然下落。另一类是在同一条件下有不确定结果的现象。例如，投掷一枚硬币，结果可能是徽面，也可能是币额面；对一次投掷来说，究竟出现哪一种结果，投掷前是无法预先确定的。又如，同一疾病的患者，服用同剂量的同一种药物后，有的痊愈，有的显效，有的无效。再如，同地区、同民族、同年龄、同性别的健康儿童，其身高值却不相同。这类现象的共同特点是：在相同条件下重复进行实验或观察时，每次结果不尽相同；一次实验或观察，究竟出现哪一种结果，事先是无法确定的。这种在同一条件下进行实验，一次实验结果不确定，而在一定数量的重复试验后呈现出统计规律性的现象，称为随机现象。随机现象在医学领域比比皆是。

医学统计学（medical statistics）是运用概率论和数理统计学的原理与方法，研究医学领域中随机现象有关数据的搜集、整理、分析与推断，进而阐明其客观规律性的一门应用科学。

医学统计学以统计描述、统计推断、关系分析以及统计设计等为主要内容，能帮助人们对其占有的信息去伪存真、由表及里。无论是基础医学、临床医学和预防医学的科学研究，还是预防、治疗、康复、保健工作的计划拟定和效果评价都离不开医学统计学。特别是计算机的问世，使多元统计分析和时间序列分析等较为复杂的统计方法得以应用，这大大增加了医学科学研究的深度和广度。

1. 医学统计学的主要内容

（1）统计研究设计　包括调查研究设计和实验研究设计。

（2）医学统计学的基本原理与方法　主要包括研究设计和数据处理中的基本统计理论和方法。例如：①资料的搜集与整理；②常用统计描述，如集中趋势与离散趋势、相对数、相关系数、回归系数、统计表与统计图等；③统计推断，如参数估计和假设检验等。

（3）医学多元统计方法　医学现象复杂多变，如疾病的发生、发展、转归、预后等受众多因素的影响。多元统计可充分利用资料的多因素信息，从而得出更贴近实际的结论。

主要方法有:多元线性回归和逐步回归分析、判别分析、聚类分析、主成分分析、因子分析、logistic 回归与 Cox 回归分析等。

(4)统计软件　计算机为大量信息的贮存与检索、复杂数据的处理提供了条件,各种统计软件的应用也日益普遍。

2.统计工作步骤　统计工作的全过程可分为4个基本步骤:设计、搜集资料、整理资料及分析资料。

(1)设计(design)　是在保证科学性、可重复性和高效性的前提下,为验证研究假说而进行的周密安排。它是在广泛查阅文献、全面了解现状、充分征询意见的基础上,在进行统计工作之前,对将要进行的研究工作所做的全面设想。它包括:明确研究目的和研究假说,确定观察对象与观察单位、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。

(2)搜集资料(collection of data)　遵循统计学原理采取必要措施得到准确可靠的原始资料。

1)搜集资料的原则　及时、准确、完整是搜集统计资料的基本原则。

2)统计资料的来源　医学领域的统计资料主要来自3个方面。一是统计报表,如医院工作报表、法定传染病报表等。统计报表是由国家统一设计,由指定的医疗卫生机构定期上报,反映居民健康和卫生服务需求状况的主要数据。对该类资料要加强漏报、重报和错报的检查。二是经常性工作记录,如疾病监测记录、健康档案、住院病历等。应注意的是该类资料具有一定的局限性,不能反映一般人群特征。三是专题调查或专题实验。对调查和实验数据,要特别注意指标说明、数据精度和编码的一并搜集等。

3)资料贮存　所搜集资料的原始记录,要妥为保存,一般在统计分析完成后也应保留一段时间。存于磁盘的资料格式,要便于汇总、再利用,并定期备份、定期复制。

(3)整理资料(sorting data)　搜集来的资料,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与归纳,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。整理资料的过程是: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;然后拟定整理表,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;最后统计归纳。

(4)分析资料(analysis of data)　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(descriptive statistics)和统计推断(inferential statistics)。前者是用统计指标、统计表、统计图等方法,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。

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。

第二节　统计学中的几个基本概念

一、同质与变异

严格地讲，同质（homogeneity）是指被研究指标的影响因素相同。但在医学研究中，有些影响因素往往是难以控制的（如遗传、营养等），甚至是未知的。因此，在实际工作中只有相对的同质。在统计学中可以把同质理解为对研究指标影响较大的、可以控制的主要因素尽可能相同。例如研究儿童的身高时，要求影响身高较大的、易控制的因素如性别、年龄、民族、地区要相同，而不易控制的遗传、营养等影响因素可以忽略。

同质基础上的各观察单位间的差异称为变异（variation）。如同性别、同年龄、同民族、同地区健康儿童的身高、体重不尽相同；相同病种、病程的病人，使用同一疗法，却未必有相同疗效。这些不同就是变异。变异是生物体的基本属性之一，也是统计研究的前提，若所研究的同质群体中各个观察单位都一样，没有差别，分析一个就够了，无须进行统计研究。

二、变　量

在搜集资料时，首先要根据研究目的确定同质观察单位，再对每个观察单位的某项特征进行测量或观察，这种特征称为变量（variable）。如上述的“身高”、“体重”、“疗效”就是变量。变量的观察结果或测量值称为变量值，变量按其值的性质可分为不同类型。

1. 数值变量（numerical variable）　其变量值是定量的，表现为数值大小，多有度量衡单位。如身高（cm）、体重（kg）、心律（次/min）、住院天数（d）、血压（mmHg 或 kPa）等。这种由数值变量的测量值组成的资料称为计量资料。大多数的数值变量为连续型变量，如身高、体重、血压等；而有的数值变量的测定值只是正整数，如心率、白细胞计数等，在医学统计学中把它们也视为连续型变量。

2. 分类变量（categorical variable）　表现为互不相容的类别或属性，亦称定性变量。分类变量可分为无序与有序两类。

无序分类变量（unordered categorical variable）是指所分类别或属性之间无程度和顺序的差别。如性别（男、女），血型（O、A、B、AB 型）等。无序分类变量的分析应先按类别分组，计各组的观察单位数，编制分类资料的频数表，所得资料称为计数资料。

有序分类变量（ordinal categorical variable）是各类别之间有程度的差别。如尿糖化验结果按 -、±、+、++、+++ 分类；疗效按治愈、好转、无效、恶化分组。有序分类变量的分析应先按等级顺序分组，计各组的观察单位数，编制各等级的频数表，所得资料称为等级资料。

变量类型不是一成不变的，可根据研究分析的需要进行转化。例如白细胞计数原属数值变量，若按正常、异常分组，则为无序分类变量；若按过低（$<4\times10^9$/L）、正常（4×10^9/L ~ 10×10^9/L）、过高（$>10\times10^9$/L）分组，则为有序分类变量。分类变量也可数量化，如可将病人的恶心反应以 0、1、2、3 表示。

数值变量的值常由测量仪器测得，受研究者主观因素影响较小，因此有较好的精确度和可重复性。分类变量的值，大多是定性描述，如患者主诉、疗效评价等，易受主观因素的影响。在临床研究中，后一种数据的重要性并不亚于前一种数据，在某些方面甚至比前者更有价值。例如临床评价某新法治疗恶性肿瘤的疗效，对患者本人来说，患者的痛感、情绪、生活自理能力等“生存质量”方面分类变量的值往往比肿瘤体积、甚至比生存时间这类数值变量的值更重要。近几年，“生存质量”评价引入临床疗效评价，并日益受到重视正是基于此因。

在做统计分析时，无论是统计描述，还是统计推断，都要先考虑变量类型，类型不同，统计方法也各异。

三、总体与样本

1. 总体(population)　总体是根据研究目的所确定的同质研究对象中所有观察单位某变量值的集合。研究目的不同，其同质含义也不同。总体具有的基本特征是同质性，即构成总体的各观察单位必须具有某种共性，这是形成总体的客观依据，也是我们确定总体范围的标准。例如对2000年某市7岁儿童体重参考值进行研究，研究对象是该市7岁健康儿童，观察单位是每个7岁健康儿童，变量是体重，变量值是体重测量值，该市2000年全体7岁健康儿童的体重值构成一个总体。它的同质基础是同地区、同年份、同为健康儿童；差异性则表现在这些儿童的体重值不相同。医学研究对象，可以是人、实验动物、微生物等；观察单位可以是一个地区、一个家庭、一个人、一只眼睛、一个细胞株、一个基因片段等。

若在某特定的时间与空间范围之内，同质研究对象的所有观察单位的某变量值的个数为有限个，则这个总体称为有限总体(finite population)。如上述7岁健康儿童体重这个研究变量，在特定的时间(2000年)，特定的空间(某市)，7岁健康儿童数是有限的，每个儿童一个体重值，所有7岁健康儿童体重值个数是有限的，这个总体为有限总体。有时总体是假设的，没有时间和空间的限制，观察单位数是无限的，称为无限总体(infinite population)。如研究碘盐对缺碘性甲状腺病的防治效果，该总体的同质基础是高发地区缺碘性甲状腺病患者，同用碘盐防治；该总体应包括已使用和设想使用碘盐防治的所有缺碘性甲状腺疾病患者，但由于设想用碘盐防治的所有缺碘性甲状腺疾病患者的防治结果是没有时间和空间限制的，因而观察单位数无限，该总体为无限总体。

2. 样本(sample)　医学研究中，有许多是无限总体，直接研究无限总体中每个观察单位是不可能的。即使是有限总体，这个“有限”也是庞大的，要对所有观察单位进行观察或研究，往往也是不可能的和不必要的。在实际工作中通常是从总体中随机抽取部分观察单位，其变量值的集合构成样本。如上例，从该市全体7岁健康儿童中随机抽取100名，他们的体重测量值构成样本。抽样研究的目的是用样本信息去推断总体特征，所以样本必须具有代表性。“代表性”是在样本来自同质总体、足够的样本含量和随机抽样的前提下实现的。所谓“随机抽样”，是指遵循随机化原则从总体中抽取样本，有多种随机抽样方法供选用。

在统计学中，描述样本变量值特征的指标称为统计量(statistics)；描述总体变量值特

征的指标称为参数(parameter)。由于个体变异的存在,即使在同一总体中随机抽取若干样本,各样本的统计量值往往不等,统计量与参数也会有所不同。这种因抽样研究所引起的差异,称为抽样误差(sampling error)。

四、概　率

1. 随机事件(random event)　对随机现象进行实验或观察称为随机试验。随机试验的各种可能结果的集合称为随机事件,简称事件。在一次随机试验中,某个随机事件可能发生、也可能不发生;但在一定数量的重复试验后该随机事件的发生情况是有规律可循的。

2. 概率(probability)　是描述随机事件发生的可能性大小的一个度量,通常用 P 表示。例如,投掷一枚均匀的硬币,随机事件 A 表示"正面向上",用 n 表示投掷次数,m 表示随机事件 A 发生的次数,f 表示随机事件 A 发生的频率 $(f=m/n)$,$0\leqslant m\leqslant n$,$0\leqslant f\leqslant 1$。用不同的投掷次数 n 作随机试验,结果如下:$m/n=8/10=0.8$,$7/20=0.35$,…,$249/500=0.498$,$501/1\ 000=0.501$,$1\ 001/2\ 000=0.500$,由此看出当投掷次数 n 足够大时,f 值稳定在 0.5,称 $P(A)=0.5$,或简写为:$P=0.5$。

当 n 足够大时,可以用 f 估计 P。对事件 A,若有 $P(A)=1$,称 A 为必然事件;若 $P(A)=0$,称 A 为不可能事件。一般,随机事件 A 的概率为 $0<P(A)<1$。

3. 小概率事件　若随机事件 A 的概率 $P(A)\leqslant\alpha$,习惯上,当 $\alpha=0.05$ 时,就称 A 为小概率事件。其统计学意义是小概率事件在一次随机试验中认为不会发生。例如,某都市大街上疾驶的汽车撞伤行人的事件的发生概率为 1/万,但大街上仍有行人,这是因为"被撞"事件是小概率事件,所以行人认为自己上街这"一次试验"中不会发生"被撞"事件。"小概率"的标准 α 是人为规定的,对于可能引起严重后果的事件,如术中大出血等,可规定 $\alpha=0.01$,甚至更小。

第二章　统计描述

统计分析包括统计描述和统计推断两部分。本章主要介绍常用的统计资料描述性指标和统计图表，以及正态分布及其应用。

第一节　数值变量资料的频数表

统计描述是用统计图（表）、统计指标来描述资料的分布规律及其数量特征的。频数表是统计描述中经常使用的基本工具之一。

一、频数表的编制

在观察值个数（即样本含量 n）较多时，为了解一组同质观察值的分布规律和便于指标的计算，可编制频数分布表，简称频数表（frequency table）。以例 2.1 说明其编制方法。

例 2.1　某地 1998 年抽样调查了 100 名 18 岁男大学生的身高（cm）资料如下，试编制频数表。

173.6	165.8	168.7	173.6	173.7	177.8	180.3	173.1	173.0	172.6
173.6	175.3	178.4	181.5	170.5	176.4	170.8	171.8	180.7	170.7
173.8	164.4	170.0	175.0	177.7	171.4	162.9	179.0	174.9	178.3
174.5	174.3	170.4	173.2	174.5	173.7	173.4	173.9	172.9	177.9
168.3	175.0	172.1	166.9	172.7	172.2	168.0	172.7	172.3	175.2
171.9	168.6	167.6	169.1	166.8	172.0	168.4	166.2	172.8	166.1
173.5	168.6	172.4	175.7	178.8	169.1	175.5	170.8	171.7	164.6
171.2	177.1	170.7	173.6	167.2	170.7	174.7	171.8	167.3	174.8
168.5	178.7	177.3	165.9	174.0	170.2	169.5	172.1	178.2	170.9
171.3	176.1	169.7	177.9	171.1	179.3	183.5	168.5	175.5	175.9

1. 求全距　找出观察值中的最大值与最小值，其差值即为全距（或极差，range），用 R 表示。本例最大值为 183.5 cm，最小值为 162.9 cm，则 $R = 183.5 - 162.9 = 20.6$（cm）。

2. 定组段和组距　根据样本含量的多少确定“组段”数，一般设 8 ~ 13 个组段。各组段的起点和终点分别称为下限和上限，某组段的组中值为该组段的（下限 + 上限）/2。相邻两组段的下限之差称为组距，常用全距的 1/10 取整做组距，以便于汇总和计算。值得注意的是，第一组段应包括全部观察值中的最小值，最末组段应包括全部观察值中的最大

值，并且同时写出其下限与上限。本例全距 20.6 的 1/10 为 2.06，取整为 2.0 cm 即组距 =2.0 cm；第一组段的下限为 162 cm，第二组段的下限为 164 cm，依次类推，最末组段为 182 ~ 184 cm，如表 2 - 1 的第(1)栏。

3. 列出频数表　把上述的组段序列制成表的形式，采用计算机或用划记法将原始数据汇总，得出各组段的观察例数，即频数，如表 2 - 1 的第(2)栏。将各组段(或各观察值)及其相应的频数列表即为频数表，如表 2 - 1 的第(1)、(2)栏。

表 2 - 1　某地 100 名 18 岁男大学生身高(cm)均数的频数表

身高组段 (1)	频数，f (2)	组中值，x (3)
162 ~	1	163
164 ~	4	165
166 ~	7	167
168 ~	12	169
170 ~	18	171
172 ~	24	173
174 ~	15	175
176 ~	8	177
178 ~	7	179
180 ~	3	181
182 ~ 184	1	183
合计	100	

二、频数分布的特征

由频数表可看出频数分布的 2 个重要特征：**集中趋势**(central tendency)和离散程度(dispersion)。例如本例，身高有高有矮，但中等身高居多，此为集中趋势；由中等身高到较矮或较高的频数分布逐渐减少，反映了离散程度。对于数值变量资料，可从集中趋势和离散程度 2 个侧面去分析其规律性。

频数分布有对称分布和偏态分布之分。对称分布是指集中位置在中央，左右两侧频数分布大致对称，如表 2 - 1 的(1)、(2)栏所示，若绘制成直方图(图 2 - 1 的 A)则更为直观清楚。偏态分布是指频数分布不对称，集中位置偏向一侧，若集中位置偏向数值小的一侧，称为正偏态分布，如表 2 - 3 的(1)、(2)栏；集中位置偏向数值大的一侧，称为负偏态分布，如冠心病、大多数恶性肿瘤等慢性病患者的年龄分布为负偏态分布。临床上正偏态分布资料较多见。不同的分布类型应选用不同的统计分析方法。

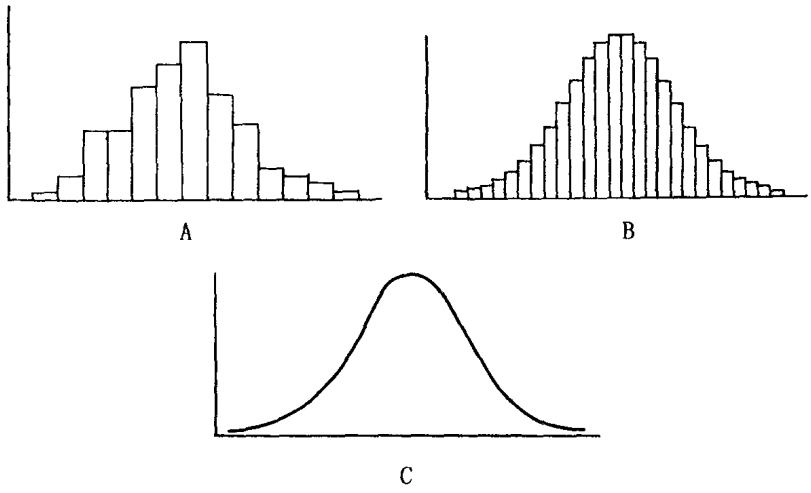

图 2－1　频数分布逐渐接近正态分布示意图

三、频数表的用途

频数表可以揭示资料分布类型和分布特征，以便选取适当的统计方法；便于进一步计算指标和统计处理；便于发现某些特大或特小的可疑值。

第二节　数值变量资料的描述性指标

一、集中趋势的描述

描述一组同质观察值的平均水平或中心位置的指标有均数、几何均数、中位数、众数、调和均数等。

（一）均数

均数（mean，average）是算术均数（arithmetic mean）的简称。常用 $\bar{X}$ 表示样本均数，μ 表示总体均数。均数用于反映一组同质观察值的平均水平，适用于正态或近似正态分布的数值变量资料。其计算方法有：

1. 直接法　用于样本含量较少时，其公式为：

$$\bar{X} = \frac{\sum X}{n} = \frac{X_1 + X_2 + \cdots X_n}{n} \tag{2.1}$$

式中，希腊字母 Σ（读作 sigma）表示求和；X_1，X_2，…，X_n 为各观察值；n 为样本含量，即观察值的个数。

例 2.2　某地 10 名 18 岁健康男大学生身高（cm）分别为 168.7，178.4，170.0，170.4，172.1，167.6，172.4，170.7，177.3，169.7，求平均身高。