第一篇　医学统计学

第一章
绪论

医学统计学是数理统计学的原理和方法在医学科学研究和医学实践中的应用。它是研究生物医学数据资料的收集、整理和分析的科学方法，对于不确定的数据通过统计分析，阐明其内在规律，做出科学推断。

作为医学工作者，学习和掌握一定的统计学知识是十分必要的。第一，在阅读医学书刊中，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含义；第二，在实际工作中，经常要做登记工作，要填写各种报表，只有懂得了原始登记与统计结果的密切关系，并掌握了收集、整理与分析资料的基本知识与技能，才能自觉地、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，做出正确的结论。

学习统计学，首先必须明确：我们掌握的关键不是数学原理，而是怎样合理地、恰当地把数理统计的方法应用到医学科研工作中去，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，重视练习与实习。做作业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。再次，各种统计符号必须写正确，汉字、阿拉伯数字与外文字母必须写清楚，不能写成模棱两可，只有在学习时养成良好的习惯，将来工作中才能少出差错。

第一节　统计学中的几个基本概念

一、变量与变量值

变量是指观察对象个体的特征或测量结果，变量又称变数，由于它在测量前不能准确预测，故常称为随机变量(random)，一般用拉丁字母表示。如患者的年龄、性别、职业等。变量的取值为变量值，如实际年龄、性别等。

二、总体与样本

总体(population)是根据研究目的所确定的同质研究对象的全体。例如研究某地区2014年10岁正常男孩身高,则该地区所有10岁男孩均为研究对象,每一个男孩为一个观察单位,全部10岁男孩的身高值构成一个总体。它的同质基础是同为某地区,同为2014年,同为10岁正常男孩。这样的总体包含有限个观察单位,故亦称为有限总体(finite population)。有时总体是假想的,例如研究10岁弱智男孩的身高,其同质基础是同为10岁男孩,同为弱智者,没有时间和空间范围的限制,因而观察单位数无限,故亦称无限总体(infinite population)。

在医学研究中,很多是无限总体,不可能直接研究总体的情况,因此经常是从总体中抽出一部分进行研究。其目的是利用较少的研究对象,花较少的人力、物力、财力来推断总体的情况。因而这一部分研究对象要求能够代表总体,这就是样本(sample),即从总体中随机抽取的一部分(有代表性的)研究对象。例如,从某地区2014年正常10岁男孩中随机抽取150名,分别测量其身高组成样本。所谓随机抽样(random sampling)的方法是指使总体中的每一个研究对象都有均等的机会被抽到样本中来的方法。样本中所包含的观察单位数在统计学中称为样本含量(sample size)或称样本例数,常用 n 表示。

三、同质和变异

同质(homogeneity)指性质相同的事物,即观察单位或观察指标受共同因素制约的部分。而在同质条件下,就同一观察指标来说,各观察单位表现出来的数量间存在着差异,这种客观存在的差异性称为变异(variation)。例如,研究儿童的身体发育,同性别、同年龄儿童(统计上称为"同质"观察单位)的身高,有高有低,各不相同,称为身高的变异。日常生活中,有时也可以看到,同性别中,有的低年龄儿童的身高,高于高年龄儿童,但总的说来,儿童的身高总是随年龄而增加,这是客观规律。同理,同种属、同性别、年龄相近的小白鼠,喂以同种饲料,其所增体重(g)亦各不相同,亦称为变异,但客观规律是饲料所含营养成分愈多,所增体重愈多。上例中的个体变异表现为定量,如身高的高低,所增体重的多少;有时亦可表现为定性的,如用某药治疗某种病人后的痊愈、好转、恶化等。同质观察单位之间的个体变异,是生物的重要特征,是偶然性的表现,是由于生物体内外环境中,多种因素的综合影响造成的,其中许多因素是未知的,也是难于控制的。统计研究的是有变异的事物,统计的任务就是在同质分组的基础上,通过对个体变异的研究,透过偶然现象,反映同质事物的本质特征和规律。

四、参数与统计量

总体的统计指标或特征值称为参数(parameter)。它是根据总体全部个体值计算的结果。参数一般是未知或假设的,但可通过样本指标进行估计。参数一般用希腊字母表示,如总体均数和总体标准差分别用 μ 和 σ 表示。统计量(statistic)是指直接从样本计算出的量数,它代表样本的特征。统计量一律用拉丁字母表示,如样本均数和标准差分别用 $\bar{X}$ 和 s 表示。

五、误差与抽样误差

测量值与真值之差为误差。根据误差产生的原因及性质可分为随机误差、系统误差和过失误差三类。随机误差是指一类不恒定、随机变化的误差，它由多种尚无法控制的因素引起。随机误差不可避免，但具有规律性。系统误差也叫规律误差，它是由于仪器结构上不够完善（如仪器不准、试剂不标准）、方法学的问题或个人问题等原因产生。其特点是受确定因素影响，观察结果成倾向性的偏大或偏小（变化有方向性但无统计规律性），通过科学设计和严格技术措施该误差可减少或消除。过失误差又称非系统误差、粗大误差或粗差。它的产生完全是由测量者（研究者）失误或测量仪器失常而造成的，这类误差应该予以避免。

由于总体中各观察单位间存在个体差异（即变异），研究中所抽取的样本，只包含总体中的一部分观察单位，因而所得样本的指标不一定恰好等于相应的总体指标。例如从某地区抽取的150名正常10岁男孩的平均身高为130.11 cm，不一定恰好等于该地正常10岁男孩的总体均数，这种由于抽样的原因所造成样本指标与总体指标之间的差异，称为抽样误差（sampling error）。显然抽样误差越小，用样本推断总体的精确度就越高。由于生物的个体变异是客观存在的，因而抽样误差是不可避免的。但是，抽样误差具有一定的规律，研究和运用这些规律正是医学统计学的重要内容。

六、频率与概率

样本的实际发生率称为频率。如在 n 次试验中，事件A发生 m 次，则比值：

$$f = m/n$$

$$= \text{A发生的实验次数/实验的总次数}$$

该式即为事件A在这n次试验中出现的频率（relative frequency）。m 为频数，频数常用小数或百分数表示，显然有：$0 \leqslant f \leqslant 1$。医学上常说的“发病率”、“患病率”和“死亡率”、“病死率”等都是指频率。

实践证明，在重复试验中，事件A的频率，随着试验次数不断增加将越来越接近一个常数 P，频数的这一特性称为频率的稳定性。频率的稳定性充分说明随机事件出现的可能是事物本身固有的一种客观属性，也是可以被认识和度量的。这个常数 P 就称为事件A出现的概率。

概率在统计学上常用“P”表示，其值在0和1之间，即 $0 \leqslant P \leqslant 1$。$P=1$ 称为必然事件，表示某事件肯定发生；$P=0$ 称为不可能事件，表示某事件肯定不发生；某些可能发生也可能不发生的事件称为随机事件或偶然事件，其概率 P 介于0和1之间，即 $0<P<1$。对于随机事件，概率越接近于1，表明其发生的可能性越大；概率越接近于0，其发生的可能性越小。统计分析的许多结论都是建立在概率大小基础上的。习惯上将 $P \leqslant 0.05$ 或 $P \leqslant 0.01$ 称为小概率事件，表示某事件发生的可能性很小。

第二节　统计资料的类型

医学统计资料按照其性质，一般分为计量资料和计数资料两大类，介于二者之间

笔记栏

的还有等级资料。不同类型的资料应采用不同的统计分析方法。

(一)计量资料

对每一个观察单位用定量的方法测定某项指标数值大小所得的资料称为计量资料,一般用度量衡单位表示。如身长(cm)、体重(kg)、血压(mmHg)、脉搏(次/min)等,都属于计量资料。每一个被研究的个体即为一个观察单位。

(二)计数资料

将观察单位按照性质或类别进行分组,然后清点各组观察单位的个数所得的资料称为计数资料。例如,调查某班学生近视眼发病情况,将学生按"近视""非近视"分组,然后清点得到每组人数等,都属于计数资料。每一个学生即为一个观察单位。

(三)等级资料

将观察单位按某种属性的不同程度分组,所得各组观察单位数,称为等级资料(ranked data)。如,一些临床生化试验,将结果按"-、±、+、…"等级分组,再清点各组所得人数;又如,临床上对治疗结果分为痊愈、显效、有效、无效、死亡等,再清点各组人数。这些均属等级资料,它既具有计数资料的性质,同时又兼有半定量性质。

根据分析需要,上述三种类型的资料可以相互转化。例如:测量某地婴儿出生体重(g),所得资料为计量资料;若按照体重是否满足 2 500 g 划分正常儿与低出生体重儿,得到各组人数为计数资料。将婴儿出生体重按其体重(g)的多少分等级:<2 500 g(低出生体重儿)、2 500 g ~(正常体重儿)、4 000 g ~(巨大儿),得到各组人数即为等级资料。

第三节　医学统计工作的基本步骤

统计工作的过程分为四个基本步骤。它们是研究设计、收集资料、整理资料和分析资料,这四个步骤是相互联系不可分割的一个整体。

一、设计

设计(design)是根据特定的研究目的,对一项医学科学研究全过程进行科学、有效和周密的计划和安排。设计主要包括专业设计和统计设计两部分内容。专业设计主要考虑专业方面的需要,如研究对象的选择,实验技术和方法的确定等。统计设计围绕专业设计确定,具体内容包括资料收集、整理和分析总的设想和安排。例如,需要收集哪方面的原始资料、用什么方式和方法收集资料、选用什么指标对收集的资料进行整理和分析以及如何控制各类误差等,可以说,科学研究的每一个步骤和各个环节都离不开设计,设计也是统计工作最关键的环节。

二、收集资料

收集资料(collection of data)是指根据实验设计的需要,完整、准确、及时地收取所需的原始数据。收集资料是统计工作的基础,收集各类资料必须坚持实事求是的原

则，以保证以后的分析结论建立在可靠的基础之上。

在生物医学领域里，统计资料可来源于统计报表，日常工作记录及专题研究三个方面。

（一）统计报表

统计报表是根据国家规定的报告制度，由医疗卫生机构定期逐级上报的一系列统计表格。卫生统计报表均由国家卫生部门与统计部门统一制订，指定有关卫生机构填报实施，例如卫生基本情况报表、法定传染病报表、职业病报表、医院工作年报表、居民病伤死亡原因年报表等。

（二）日常工作记录

医务工作者日常的工作记录都可以作为统计资料的来源，如门诊登记、住院病历、体检记录、出生、死亡、肿瘤登记等。因而，在日常工作中，医务人员应根据表格项目和填报要求，认真填写上述各种工作记录，从而保证统计资料的完整、准确。

（三）专题调查或实验

这是根据医学研究的目的，临时组织进行的专业性调查。当上述两方面不能满足需要时，经常采用这种方法。例如，某地10岁男孩发育状况的调查研究、全国千分之一人口生育率调查、全国糖尿病调查、大蒜对煤焦沥青致癌阻断作用的实验研究等。

三、整理资料

根据设计要求，对所收集到的资料进行分类、归纳、汇总的方法称为整理资料（sorting data）。整理使资料系统化、条理化、便于进一步的计算分析。

（一）核查资料

检查所收集到的原始数据，确定其是否完整，有无遗漏，记录是否正确、各项目数据间有无逻辑错误。发现问题应及时更正、补充。对无法进行补救的数据或项目应将其剔除。

（二）设计分组

根据资料的性质或数量特征进行分组、整理，以反映其内在的规律。分组可按照研究事物的性质或数量进行划分。

1. 按性质分组　即根据事物的种类或属性进行分组。如按照人群性别、职业进行分组；按照疾病疗效的痊愈、好转、无效、死亡分组或按照化验结果的阳性、阴性分组等。

2. 按数量进行分组　即根据测量数据量的大小进行分组。如按照人群的年龄、身高、体重进行分组；按照治疗前后体内血红蛋白的量或微量元素的量进行分组等。

在进行量的分组时应注意划分适当组数。分组过小，误差较大，往往是掩盖了事物的特点。例如，研究近视眼的发病情况，把0～20岁划为一组，则很难发现青少年时期的发病高峰特点；分组过多则不容易突出事物的规律性。因此，当还不了解研究现象的内在规律时，可先分组细一些，根据情况再适当并组，分粗一些。一般来说，分组以8～15组为宜。

（三）拟整理表

整理表（sorting table）是作为归纳汇总用的一种表格，根据研究目的和要求，把关系密切的几个项目制订在一个表内，使表达其相互关系和固有规律，表格的制作要求，详见统计图表一章。

（四）归纳汇总

按整理表内容将原始数据归纳计数的过程为归纳汇总。常用的方法有画记法和分卡法等。

1. 画记法　画记法（tabulation method）也叫列表法，即用画“正”的方法来计数（详见下章频数表的编制方法）。这种方法简便，但出错后不易检查，故适用于项目较少或数据量不大的资料。

2. 分卡法　分卡法（card-sorting method）是将登记卡根据分组设计，分别置于相应组的位置进行归类，然后清点计数。例如，将化验结果按某种标志分为五组，研究对象每人一卡，按其化验结果分别将卡片放置于不同位置，然后清点每种结果卡片的数量。此方法比较简便，易于检查改正错误，但成本较高，主要适于数量较少的卡片资料。

上述两种手工归纳汇总方法是较常用的实用有效的方法，目前由于电子计算机的发展和普及，用计算机对数据进行整理、归纳、计数的方法将逐渐取代手工方法。

四、分析资料

分析资料（analysis of data）是统计工作的最后步骤。根据设计要求，计算有关指标，并进行统计描述和统计推断，结合专业知识、阐明事物的内在规律。将统计结果用统计表和统计图加以科学说明，最后写出文字总结。如何进行统计资料的分析，将是我们医学统计方法重点阐述的内容。

思考练习题

1. 统计工作的基本步骤是什么？
2. 统计资料可分为几类，各有何特点？
3. 何谓抽样误差？抽样误差能否避免？为什么？

第二章
数值变量资料的统计分析

第一节　平均数与标准差

数值变量资料又称计量资料，是指用定量方法测定观察单位某项指标数值大小所得的资料。对于数值变量资料，最常用的统计描述指标是平均数和标准差。平均数表示一组同质观察值（数据）的平均水平或集中趋势；标准差表示其变异程度或离散趋势。

一、频数分布表

计量资料的观察值，可用频数分布表（频数表，frequency table）来描述其分布规律。

（一）频数表的编制

频数是指在一个抽样资料中某变量值出现的次数。将各变量值及其频数列表即为频数表。现以某市男童身高资料为例，按以下步骤编制频数表。

例2-1　某市2016年110名7岁男童身高（cm）资料如下（表2-1），试编制频数表。

表2-1　某市2016年110名7岁男童身高（cm）资料

112.4	117.2	122.7	123.0	113.0	110.8	118.2	108.2	118.9	118.1
123.5	118.3	120.3	116.2	114.7	119.7	114.8	119.6	113.2	120.0
119.7	116.8	119.8	122.5	119.7	120.7	114.3	122.0	117.0	122.5
119.8	122.9	128.0	121.5	126.1	117.7	124.1	129.3	121.8	112.7
120.2	120.8	126.6	120.0	130.5	120.0	121.5	114.3	124.1	117.2
124.4	116.4	119.0	117.1	114.9	129.1	118.4	113.2	116.0	120.4
112.3	114.9	124.4	112.2	125.2	116.3	125.8	121.0	115.4	121.2
123.8	120.1	118.4	122.8	120.1	112.4	118.5	113.0	120.8	114.8
117.9	119.1	122.8	120.7	117.4	126.2	122.1	125.2	118.0	120.7
116.3	125.1	120.5	114.3	123.1	122.4	110.3	119.3	125.0	111.5
116.8	125.6	123.2	119.5	120.5	127.1	120.6	132.5	116.3	130.8

1. 求全距　全距(range)即最大值与最小值之差,用 R 表示。本例最大值为132.5 cm,最小值为108.2 cm,则 $R=132.5-108.2=24.3$(cm)。

2. 定组距　频数表组段的划分是以统计分析和统计计算为目的,过多过少均不适宜,一般取8~15组为宜。组与组之间的差值称组距(class interval),用 i 表示。组距可以相等或不等,一般取相等的组距。组距 $i=R/$预分组段数。本例若取预分组段数为10,则 $i=24.3/10=2.43$(cm)。为方便计算,取整数2.0 cm作为组距。

3. 划分组段　每一个组段都有一个起始值作为下限(lower limit)和一个终止值作为上限(higher limit)。分组时,第一组应取在最小值附近并包括最小值的整数,而最后一组则应包括最大值。为了避免两组界限交互包含,组段常用各组的下限及"~"表示,最后一组则应包括其上限值(封口)。如:

第一组　108~
第二组　110~
⋮　　⋮
最末组　132~134

4. 列表画记　对所划分组组段进行列表,按画记的方法用"正"字号对原始资料进行计数,得到各组段观察值的频数,如表2-2所示。

表2-2　110名7岁男童身高的画记

身高/cm	画记	频数
108~	丨	1
110~	下	3
112~	正丅	9
114~	正丅	9
116~	正正正	15
118~	正正正下	18
120~	正正正正一	21
122~	正正丅	14
124~	正正	10
126~	丅	4
128~	下	3
130~	丅	2
132~134	一	1
合计	—	110

(二)频数表的用途

1. 揭示频数分布的特征　从表2-1可以看出,7岁男童身高值有高有低,但过高过低的人数均较少,而居中者较多。频数集中在"120 cm~"组段,反映了集中趋势;

最低 108 cm，最高 134 cm，这种参差不齐反映了离散趋势。集中趋势和离散趋势是频数分布的两个重要特征。

2. 揭示频数分布的类型　从频数表可以看出频数分布是否对称。若频数分布基本对称，且中间频数分布较多，则大体上可以视为正态分布；若频数分布不对称，集中位置偏向一侧，则为偏态分布。由此便于决定进一步的统计计算和统计分析方法的选择。

3. 便于发现某些特大或特小的可疑值　从频数分布表可以发现一些可疑的特大或特小值，以便进一步检查后决定其取舍。

二、平均数

描述一组同质变量值平均水平或集中趋势的指标是平均数（average）。平均数在统计学中应用非常广泛。根据数据分布类型的不同，平均数有不同的算法，常用的平均数有算术平均数、几何均数和中位数。

（一）算术平均数

算术平均数（arithmetic mean）简称均数（mean），反映一组分布均匀（近似正态分布）的变量值的平均水平。常用符号 $\overline{X}$ 表示样本均数，用符号 μ 表示总体均数。均数的计算可用直接法或间接法进行。间接法计算中又可使用加权法或简化法进行计算。这里只介绍直接法和加权法。

1. 直接法　适用于样本含量较小，即变量个数较少的未分组资料。计算时将所有观察值 X_1、X_2、……、X_n 直接相加再除以样本例数 n。

$$\overline{X}=\frac{X_1+X_2+X_3+X_4+\cdots X_n}{n}=\frac{\sum X}{n} \tag{2-1}$$

式中 $\sum$ 是希腊字母，读作 Sigma，为求和符号。

例 2-2　10 名 5 岁男童身高值（cm）分别为：96，98，100，101，102，105，110，100，103，104。求其平均身高。

应用式 2-1，得：

$$\overline{X}=\frac{96+98+100+101+102+105+110+100+103+104}{10}=\frac{1019}{10}=101.9(\text{cm})$$

2. 加权法　加权法适用于样本含量较大的分组资料。计算时先对原始资料列加权法计算表，求出各组段频数及组中值。组中值即为每一组段的中间值，可用本组段下限+下组段下限/2 算得。然后再以各组段组中值（X）与频数（f）乘积之和（$\sum fX$）除以总频数（$\sum f$），即样本含量（n）。

$$\overline{X}=\frac{f_1X_1+f_2X_2+f_3X_3+\cdots+f_nX_n}{f_1+f_2+f_3+\cdots+f_n}=\frac{\sum fX}{\sum f} \tag{2-2}$$

式中：f 为频数，X 为各组段组中值，$\sum f$ 为总例数（即 n），$\sum fX$ 为各组段组中值与频数乘积之和。

式中各组段的频数起了“权数”的作用，它权衡了各组中值由于频数不同对均数

的影响，故本法称为加权法。

例 2-3　对表 2-1 资料用加权法计算平均身高。

首先列加权法计算表（表 2-3）：

$$\overline{X}=\frac{1\times109+3\times111+9\times113+\cdots+1\times133}{1+2+3+\cdots+1}=\frac{13\ 194}{110}=119.95(\text{cm})$$

7 岁男童平均身高为 119.95 cm。

表 2-3　110 名 7 岁男童身高均数的加权法计算表

身高组段	频数(f)	组中值(X)	fX
(1)	(2)	(3)	(4)=(2)×(3)
108 ~	1	109	109
110 ~	3	111	333
112 ~	9	113	1 017
114 ~	9	115	1 035
116 ~	15	117	1 755
118 ~	18	119	2 142
120 ~	21	121	2 541
122 ~	14	123	1 722
124 ~	10	125	1 250
126 ~	4	127	508
128 ~	3	129	387
130 ~	2	131	262
132 ~ 134	1	133	133
合计	110($\sum f$)	—	13 194($\sum fX$)

（二）几何均数

当一组观察值之间是倍数关系（即呈对数正态分布的偏态资料 ）时，如抗体的滴度；某些传染病的潜伏期等，可用几何均数（geomitric mean）表示其平均水平，用符号 G 表示。

当样本含量较小时，可用直接法计算几何均数。计算公式：

$$G=\sqrt[n]{X_1X_2X_3\cdots\cdots X_n}$$

或

$$G=\lg^{-1}(\sum \lg X/n) \qquad (2-3)$$

例 2-4　设有 5 份血清的抗体效价为 1∶10，1∶100，1∶1 000，1∶10 000，1∶100 000，求其平均效价。利用血清抗体效价倒数值代入公式 2-3，得：

$$\overline{G}=\overline{\lg^{-1}}\left(\frac{\lg10+\lg100+\lg1000+\lg10000+\lg100000}{5}\right)=\lg^{-1}3=1\ 000$$

血清抗体平均效价为 1∶1 000。

当变量值为个数较多的频数分组资料时，可用加权法进行计算。计算公式：

$$G = \lg^{-1}\left(\frac{\sum f\lg x}{\sum f}\right) \tag{2-4}$$

例2-5　某医院研究HBsAg阳性的慢性迁延性肝炎患者（简称HBsAg+CPH）的HBsAg滴度资料见表2-4第（1）、（2）栏，求其平均滴度。

表2-4　HBsAg(+)CPH患者HBsAg滴度几何均数计算表

抗体滴度	人数(f)	滴度倒数(X)	lgx	flgx
(1)	(2)	(3)	(4)	(5)=(2)×(4)
1∶16	6	16	1.204 12	7.224 72
1∶32	12	32	1.505 15	18.061 80
1∶64	10	64	1.806 18	18.061 80
1∶128	6	128	2.107 21	12.643 26
1∶256	10	256	2.408 24	24.082 40
1∶512	7	512	2.709 27	18.964 89
1∶1024	4	1024	3.010 30	12.041 20
	$\sum f=55$			$\sum f\lg x=111.08007$

$$\lg G = \frac{\sum f\lg x}{\sum f} = \frac{111.08007}{55} = 2.01964$$

$$G = \lg^{-1}2.01964 = 104.6$$

即这55例HBsAg(+)CPH患者HBsAg的平均滴度为1∶105

在计算几何均数时注意：①观察值不能有0，因为0与任何数无倍数关系，也不能取对数值。②观察值不能同时有正有负，因为负数无对数。若观察值全为负数，可先把负号去掉，计算后再加上。

（三）中位数和百分位数

均数和几何均数的运用均要求资料是一定的分布状态。实际工作中某些资料的分布并不知道，或有的资料呈偏态分布；或有的资料会出现特大或特小数值；甚至资料一端或两端无确定数值时，均应选用中位数（median）来代表其平均水平。百分位数（percentile）是一种位置指标，以Px表示。一个百分位数将总体或样本的全部观察值分为两部分，理论上有$X\%$的观察值比它小，有$(100-X)\%$的观察值比它大。

中位数是一个特定的百分位数，即P_{50}，在全部观察值中，有一半比它大，有一半比它小，它是全部观察值按大小顺序排列，位次居中的数值，在所有的百分位数中，应用最广，特用M表示。

（1）小样本资料中位数计算　先将观察值按大小顺序排列，再按下式计算。

当n为奇数时　　$M = X_{(\frac{n+1}{2})}$　　(2-5)

当 n 为偶数时 $$M=\frac{\left[X\left(\frac{n}{2}\right)+X\left(\frac{n}{2}+1\right)\right]}{2} \quad (2-6)$$

式中 n 为一组观察值的总个数，$\frac{(n+1)}{2}$、$\frac{n}{2}$、$\left(\frac{n}{2}+1\right)$ 均为观察值的位次。

例 2-6　某传染病患者 5 人，其潜伏期(d)分别为 2,3,5,8,14，求其中位数？

$n=5$，为奇数，代入公式 2-5，得：

$M=X=X_3=5$(d)

8 名新生儿身长(cm)依次为 50,51,52,53,54,54,55,58。求其中位数。

n=8，为偶数，代入公式 2-6，即可求得。

对于大样本资料，可采用频数表法(分组法)。

(2)大样本资料中位数和百分位数计算　步骤是：①按所分组段，由小到大计算累计频数和累计频率。②按下式求中位数 M 或其他百分位数 P_x。计算公式为：

$$P_x=L+\frac{i}{f_x}(nx\%-f_L) \quad (2-7)$$

式中 L 为百分位数所在组段的下限，i 为百分位数所在组段的组距，f_x 为百分位数所在组的频数，n 为总频数，P_x 为百位数，f_L 为比 L 小的累计频数。

例 2-7　某市大气中 SO_2 的日平均浓度见表 2-5 第(1)、(2)栏，分别求第 25 百分位数及中位数。

表 2-5　某市大气中 SO_2 日平均浓度的中位数及百分位数的计算

浓度($\mu g/m^3$)	天数(f)	累计频数	累计频率(%)
(1)	(2)	(3)	(4)
25 ~	39	39	10.8
50 ~	67	106	29.4
75 ~	64	170	47.1
100 ~	63	233	64.5
125 ~	45	278	77.0
150 ~	30	308	85.3
175 ~	17	325	90.0
200 ~	9	334	92.5
225 ~	7	341	94.5
250 ~	6	347	96.1
275 ~	5	352	97.5
300 ~	3	355	98.3
325 ~	6	361	100.0

计算累计频数，见第(3)栏，以及累计频率，见第(4)栏，本例 $n=361$。

求 P_{25} 由(1)、(4)栏可见，25% 在 10.8% 与 29.4% 之间，故 P_{25} 在“50 ~”组段内。今 L=50，$i=25$，$f_{25}=67$，$f_L=39$，代入式 4-8，得

$$P_{25}=50+25/67(361\times25\%-39)=69.12(g/m^3)$$

求 M　因 $M=P_{50}$，同理知 M 在“100～”组段内，L=100，$i=25$，$f_{50}=63$，$f_L=170$ 代入式 2-7 得

$$M=100+\frac{25}{63}(361/2-170)=104.17(g/m^3)$$

式中 L、i_M、f_M 分别为中位数所在组段的下限、组距和频数，f_L 为中位数所在组之前各组段的累积频数。

三、标准差

均数作为一代表值说明一组资料的平均水平，但没有说明各个个体之间的差异。例如：同某地 7 岁男童，有的身材高一些，有的身材矮一些，而他们都是正常身高的男孩。这种个体间的差异是由遗传和环境诸方面因素造成的。

（一）变异指标

表示变异的指标有很多，常用的有全距、方差和标准差。而其中又以标准差最为常用。

1. 全距　全距（range）又称极差，为一组数据中最大值与最小值之差。极差大说明变异程度大，极差小说明变异程度小。

2. 方差　方差（variance）是指总体中每一个变量值与均数之差平方和（亦称离均差平方和）的均数。它考虑到组内每一个变量值（X）与均数（μ）距离的大小对变异程度的影响。若以 σ^2 表示总体方差，则表达式为：

$$\sigma^2=\frac{\sum(X-\mu)^2}{n} \tag{2-8}$$

3. 标准差　标准差（standard deviation）为方差的平方根。因方差的单位是变量值单位的平方，不便与原来的数据进行比较，故采用标准差。标准差以符号 σ 表示，其计算式为：

$$\sigma=\frac{\sum(X-\mu)^2}{n} \tag{2-9}$$

式中：X 为变量值，μ 为总体均数，n 为样本含量。

由于在实际应用中，μ 是未知数，因而常用样本均数 $\bar{X}$ 代替 μ，得到样本标准差 S。计算公式：

$$S=\frac{\sum(X-\bar{X})^2}{n-1} \tag{2-10}$$

式中以 $n-1$ 代替 N，是因为在抽样研究中抽样误差的存在，导致 $\bar{X}$ 常不等于 μ，而使结果偏小。数理研究证明：$\sum(X-\bar{X})^2<\sum(X-\mu)^2$，因而使用 $n-1$ 作分母，使得样本标准差 S 值更接近于总体标准差 μ 值。

数理统计研究结果表明：$\sum(X-\bar{X})^2=\sum X^2-(\sum X)^2$，故式 2-10 可写成：

$$S=\sqrt{\frac{\sum(X-\bar{X})^2}{n-1}}=\sqrt{\frac{\sum X^2-(\sum X)^2/n}{n-1}} \tag{2-11}$$

综上所述，由标准差来反映一组观察值的变异程度比全距和方差都要完善，因而能够得到广泛的应用。标准差愈小，说明变量值的变异程度也愈小；标准差愈大，说明变异程度也愈大。

（二）标准差的计算

1. 直接法　适用于样本含量较小的未分组资料。

例 2-8　10 名男婴出生体重（kg）分别为：2.65，2.38，2.65，3.15，3.55，2.40，4.00，3.00，2.83，3.30。求标准差。

列标准差计算，如表 2-6。

表 2-6　10 名男婴出生体重标准差

男婴号	X	X^2
1	2.65	7.0225
2	2.38	5.6644
3	2.65	7.0225
4	3.15	9.9225
5	3.55	12.6025
6	2.40	5.7600
7	4.00	16.0000
8	3.00	9.0000
9	2.83	8.0089
10	3.30	10.8900
合计	29.91（$\sum X$）	91.8933（$\sum X^2$）

代入公式 2-11，得：

$$S=\sqrt{\frac{\sum(X-\bar{X})^2}{n-1}}=\sqrt{\frac{\sum X^2-(\sum X)^2/n}{n-1}}=0.52(\text{kg})$$

10 名男婴出生体重标准差为 0.52 kg。

2. 加权法　适用于样本含量较大的分组资料，其计算公式为：

$$S=\sqrt{\frac{\sum f_i x_i^2=z-\frac{(\sum f_i x_i)^2}{\sum f_i}}{\sum f_i-1}} \qquad (2-12)$$

式中：X 为组中值，f 为频数。

例 2-9　仍以表 2-1 资料为例，求 110 名 7 岁男童身高标准差？

列加权法计算表，如表 2-7：

表 2-7　110 名 7 岁男童身高标准差加权法计算表

身高组段	频数(f)	组中值(X)	fX	fX^2
(1)	(2)	(3)	(4)=(2)×(3)	(5)=(3)×(4)
108 ~	1	109	109	11 881
110 ~	3	111	333	36 963
112 ~	9	113	1 017	114 921
114 ~	9	115	1 035	119 025
116 ~	15	117	1 755	205 335
118 ~	18	119	2 142	254 898
120 ~	21	121	2 541	307 461
122 ~	14	123	1 722	211 806
124 ~	10	125	1 250	156 250
126 ~	4	127	508	64 516
128 ~	3	129	387	49 923
130 ~	2	131	262	34 322
132 ~ 134	1	133	133	17 689
合计		110($\sum f$)—	13 194($\sum fX$)	1584 990($\sum fX^2$)

代入式 2-12，得：$S=4.72$(cm)；110 名 7 岁男童身高标准差为 4.72 cm。

(三)标准差的用途

标准差是医学上最常用的表示变异程度的指标，其主要用途有以下几点：

1. 表示变量值的变异程度或离散程度　当两组资料均数相近、单位相同时，标准差大，说明变异程度大，数值较分散，平均数的代表性较差；反之，标准差小，说明变异程度小，数值较集中，平均数的代表性较好。

2. 描述规律　结合均数，描述正态曲线下面积的分布规律。

3. 计算变异系数　变异系数(coefficient of variation)指同一组变量值标准差与均数的百分比，符号为 CV，主要用于比较单位不同或单位相同，但均数之间相差较悬殊的几组资料之间的变异程度。计算公式为：

$$CV=\frac{S}{\overline{X}}=100\% \tag{2-13}$$

例 2-10　某地 10 名男婴，出生体重的平均数为 2.99 kg，标准差为 0.52 kg；身高的平均数为 52.25 cm，标准差为 6.65 cm，比较两者的变异程度。

体重：$CV=\frac{0.52}{2.99}\times 100\%=17.39\%$

身高：$CV=\frac{6.65}{52.25}\times 100\%=12.73\%$

可见 10 名男婴体重的变异程度大于身高，或说身高值较体重值稳定。

4. 计算标准误　详见下节。

四、正态分布及其应用

(一)正态分布

若将表 2-1 的频数表资料绘制成直方图;并设想观察人数不断增多;组段不断分细;将变窄直条的中点连接起来,形成一条光滑曲线,如图 2-1 所示。

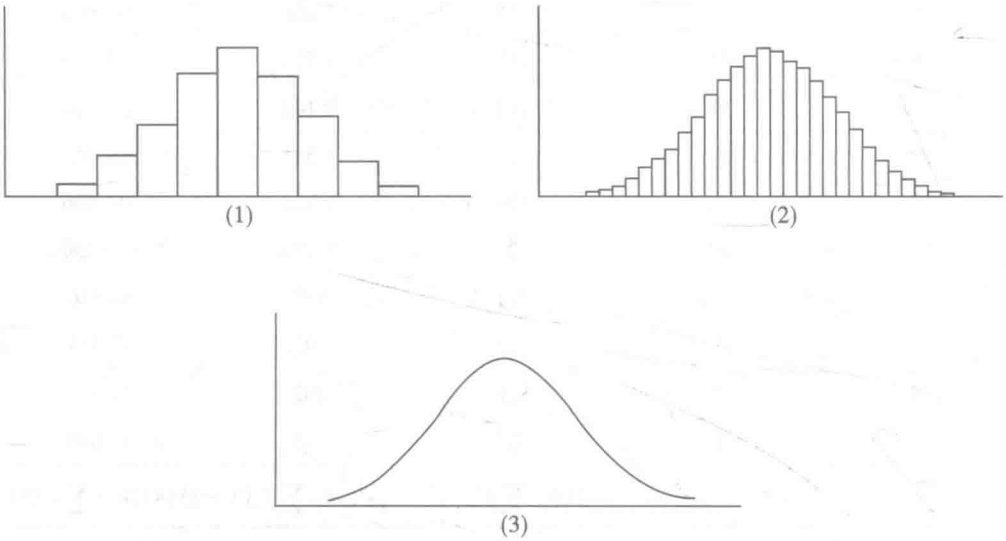

图 2-1　频数分布逐渐接近正态分布

这种以均数为中心,靠近均数的位置频数分布较多,远离均数的位置频数分布较少,且左右对称的频数分布称为正态分布(normal distribution)亦称 Gauss 分布。在医学和生物学领域中,许多数据资料的分布(如身高、血压、血糖等)均呈正态分布或近似正态分布。所以正态分布是一种最重要、最常见的连续型分布。正态分布具有以下特征:

(1)对称性　正态分布以均数为中心,左右对称。

(2)集中性　正态曲线在横轴上方均数处最高。

(3)均匀变动性　均数两侧频数的减少是逐渐和均匀的。

(4)正态分布有两个参数　即均数和标准差。均数是位置参数,标准差为形态参数。

为了方便应用,常将数据作变量变换,以均数为 0,标准差为单位,得到正态离差值 U。

$$U=\frac{\overline{X}-\mu}{\sigma} \tag{2-14}$$

U 值的分布也为正态分布,且称之为标准正态分布(standard normal distribution)。研究和应用标准正态分布的规律,也是统计学重要内容之一,它是许多统计方法建立的基础。

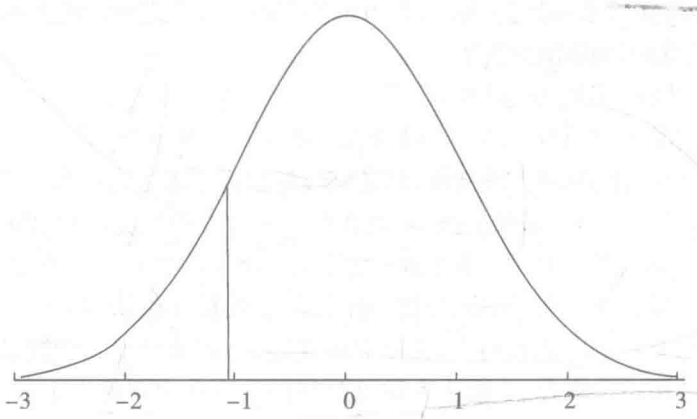

图 2-2　标准正态分布

(二)正态曲线下面积的分布规律

正态分布曲线简称正态曲线(normal curve),是一条高峰位于中央,两侧逐渐下降并完全对称,两端永不与横轴相交的钟形曲线。数理统计研究表明:横轴以上、正态曲线以下面积的分布是有一定规律的。在统计学上,以下几条规律特别重要,由图 2-3 说明。

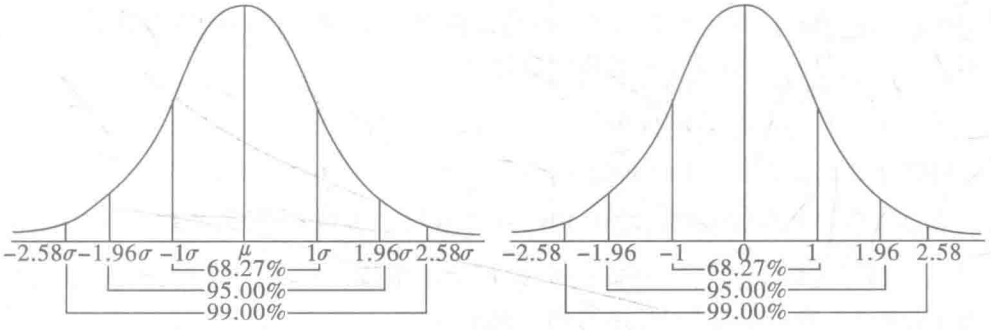

图 2-3　正态曲线及面积分布

图中以 μ 表示总体均数,σ 为总体标准差,以曲线下总面积为 1(或 100%),则正态曲线下面积的分布有以下(特别)规律:

(1)$\mu \pm \sigma$ 范围内,曲线下面积占总面积的 68.27%

(2)$\mu \pm 1.96\sigma$ 范围内,曲线下面积占总面积的 95.00%。

(3)$\mu \pm 2.58\sigma$ 范围内,曲线下面积占总面积的 99.00%。

也就是说,在上述曲线范围内,分别有 68.27%,95.00% 和 99.00% 的变量值分布在其中。在医学研究中,正态分布的规律非常有用,常常以此确定临床正常值范围。

(三)正常值范围的制定。

正常人的解剖、生理、生化等各种生理常数称为正常值(normal value),由于个体间变异的存在,这些生理常数在正常人群中有一定的波动范围,而医务人员经常需要此范围作为判断正常与异常界限的参考值。例如,正常人白细胞数参考值为 4 ~ 10×10^9/L,正常成年女性血红蛋白量为 110 ~ 150 g/L 等。这些正常值范围可利用正态曲