高等医药院校选修教材

供基础、临床、口腔医学类专业用

医学统计学

倪宗瓒 主编

人民卫生出版社

高等医药院校选修教材

（供基础、临床、口腔医学类专业用）

医 学 统 计 学

倪宗瓒　主编

倪宗瓒（华西医科大学）
余松林（同济医科大学）
杨　珉（华西医科大学）　　编写
詹绍康（上海医科大学）

人 民 卫 生 出 版 社

这套教材原为卫生部组织的统编教材，选经修订，现改为卫生部推荐教材，供各地院校选用。全套教材共45种，其中必修课教材37种，选修课教材8种，均经卫生部聘任的高等医学院校临床医学专业教材评审委员会审定。

必修课教材

1.《医用高等数学》	胡纪湘 主 编
2.《医用物理学》第三版	邝华俊 主 编
3.《基础化学》第三版	丁绪亮 主 编
4.《有机化学》第三版	徐景达 主 编
5.《医用生物学》第三版	李 璞 主 编
6.《系统解剖学》第三版	郑思竟 主 编
7.《局部解剖学》第三版	徐恩多 主 编
8.《解剖学》	王永贵 主 编
9 《组织学与胚胎学》第三版	成令忠 主 编
10.《生物化学》第三版	顾天爵 主 编
11.《生理学》第三版	周衍椒 主 编
	张镜如 副主编
12.《医用微生物学》第三版	陆德源 主 编
13.《人体寄生虫学》第三版	徐秉锟 主 编
14.《医学疫免学》	郑武飞 主 编
15.《病理学》第三版	武忠弼 主 编
16.《病理生理学》第三版	冯新为 主 编
17.《药理学》第三版	江明性 主 编
18.《医学心理学》	李心天 主 编
19.《法医学》第二版	郭景元 主 编
20.《诊断学》第三版	戚仁铎 主 编
21.《放射诊断学》第二版	吴惠恩 主 编
22.《内科学》第三版	陈灏珠 主 编
	李宗明 副主编
23.《外科学》第三版	裘法祖 主 编
	孟承伟 副主编
24.《妇产科学》第三版	郑怀美 主 编
25.《儿科学》第三版	左启华 主 编
26.《神经病学》第二版	黄友歧 主 编
27.《精神病学》第二版	沈渔邨 主 编
28.《传染病学》第三版	彭文伟 主 编
29.《眼科学》第三版	毛文书 主 编
	孙信孚 副主编
30.《耳鼻咽喉科学》第三版	黄选兆 主 编
31.《口腔科学》第三版	毛祖彝 主 编
32.《皮肤性病学》第三版	王光超 主 编
33.《核医学》第三版	周 申 主 编
34.《流行病学》第三版	耿贯一 主 编
35.《卫生学》第三版	王翔朴 主 编
36.《预防医学》	顾学箕 主 编
37.《中医学》第三版	贺志光 主 编

选修课教材

38.《医学物理学》	刘普和 主 编
39.《医用电子学》	刘 骥 主 编
40.《电子计算机基础》	华蕴博 主 编
41.《医学遗传学基础》	杜传书 主 编
42.《临床药理学》	徐叔云 主 编
43.《医学统计学》	倪宗瓒 主 编
44.《医德学概论》	丘祥兴 主 编
45.《医学辩证法》	彭瑞骢 主 编

以上教材均由人民卫生出版社出版，新华书店总店科技发行所发行。

全国高等医学院校临床医学专业第二届教材评审委员会

主任委员　裘法祖

副主任委员　高贤华

委　员（以姓氏笔画为序）

方　圻　毛文书　刘士杰　刘湘云　乔健天　沈渔邨

武忠弼　苏应宽　金有豫　南　溯　胡纪湘　顾天爵

编写说明

本书是根据1987年5月高等医药院校临床医学专业教材编审工作会议提出的，在原有《卫生学》中医学统计方法的基础上，增设一门《医学统计学》选修课教材的要求而编写的。

全书共十二章。第一章讲述基础统计知识，除简要回顾了医学统计方法中的基本统计原理和方法外，还对其作了必要的延伸和补充。第二章讲述临床试验设计。第三至十二章介绍临床试验中常用的医学统计方法。其中记有"*"号的章节为选学内容。书中共有两个附录，包括与本书内容有关的统计用表和实习题，后者供课间实习选题参考。

本书的重点是介绍临床试验设计的原理和方法，编写中并注意吸收临床流行病学中有关医学实验设计的观点。由于调查设计和实验设计部分已在《卫生学》一书中作了介绍，这里不再赘述。

近年来，医学统计学发展十分迅速，电子计算机的应用和普及使得计算繁杂的统计方法能应用于医学科学研究的实践中。因此，本书尽可能较全面地介绍医学统计方法。对于某些计算复杂的统计方法，例如多元线性回归，判别分析，聚类分析，logistic 回归和Cox 模型等，书中也作了简略的介绍，以便于学生在今后的医学实践中分析应用。

本书在讲述医学统计学的基本原理、基本概念、基本方法，以及各种方法的注意事项和适用条件时，尽量采用从实例入手，便于学生实际应用。

本书编写过程中，华西医科大学的祝绍琪教授参加了定稿会，并对相关与回归和多因素分析等部分章节进行了审阅和修改。杨树勤教授对本书的编写方法和书中的内容选择进行了具体指导。本书的学术秘书倪旱雨和刘继兰同志在编写过程中作了大量工作。潘晓平、张俊同志负责校核了书中的计算部分。对于他们在本书编写过程中付出的辛勤劳动，谨此致谢。

限于编者的水平和经验，虽然我们力图提高质量，但一定存在不少缺点和错误，希望使用本教材的师生批评指正，以利改正。

倪宗瓒

目　录

第一章　基础统计方法

1.1　绪　　言

1．医学统计学在医学科学中的地位和作用　医学统计学（medical statistics）是应用概率论和数理统计的基本原理和方法，结合医学实际，研究资料和信息的搜集、整理与分析的一门学科。近代医学发展十分迅速，许多新的问题需要人们去研究解决，认识其内在的联系。医学统计学正是一门帮助人们透过许多偶然现象分析和判断事物的内在规律的科学。电子计算机技术的普及和发展，为大量统计资料和信息的贮存、整理和分析，提供了有利的条件。许多供医学统计设计和整理分析专用的统计程序，既便于医务工作者应用医学统计方法解决医学科学中的实际问题，又增加了应用一些复杂的统计方法（如多变量分析等）进行医学科学研究的可行性。因此，医学统计学已成为促进医学发展的一门重要的应用科学，成为医学生应具有的一种分析和解决问题的重要手段。

近年来，在发达国家中兴起了对医务工作者，特别是临床医师进行继续教育的培训计划，称为D．M．E.即设计、测量和评价（design，measurement and evaluation）。它的主要内容是应用医学统计的思维和分析方法，结合其它相关学科，引导人们去正确地阅读文献资料，进行日常工作，开展医学科学研究，总结工作经验等。这门学科已受到医学界的重视，越来越多的医务工作者认识到它在医学科学中的地位和作用，这使医学统计学的应用领域更加广阔，联系医学实际更为密切。

2．医学统计学的主要内容　医学统计学的主要内容是研究医学统计设计、数理统计方法在医学科学中的应用。根据目前医学生的现状，本书着重介绍以下内容：

（1）医学统计研究设计　进行医学科研设计时，除应用必要的专业知识外，必须应用医学统计设计的基本原理进行周密的考虑，采取必要的有效措施以保证研究的结果能够回答研究假设中提出的问题，使用较少的人力、物力和时间以取得较好的效果。

医学科研设计可按照是否对研究对象进行干预，分为调查设计和实验设计两大类：调查（survey）是客观地反映事物的实际情况，未加任何干预措施，为进一步决策和深入研究提供依据。如为了解1988年某省恶性肿瘤死亡率的分布；某地学龄前儿童贫血的患病率等。实验研究中研究者可根据研究目的主动加以干预措施，控制非试验因素的干扰，并观察总结其结果，回答研究假设所提出的问题。例如研究魔芋精粉可否降低大白鼠血中的胆固醇含量，首先可以将若干条件相近的大白鼠，随机地分配入两组，用高胆固醇饲料和高胆固醇加魔芋精粉饲料作为干预措施，经过观察总结和分析，得出魔芋精粉是否有降低胆固醇作用的结论。由此可见，实验研究与调查研究不同之处在于：实验研究中研究者主动加入了干预措施，并根据研究目的加以必要的控制条件。

实验研究设计根据研究对象不同又可分为：以实验动物和实验样品为对象的实验设计和以人为对象的临床试验设计。前者可以根据研究目的采用较严格的控制措施；后者在设计时必须充分考虑以人为对象的特点。临床试验设计在医学科研实践中较为常用，设计的基本原理和方法也适用于实验设计，故本书将作详细介绍（见第二章）。

(2) 常用的基本统计方法　包括①定量和定性资料的统计描述和总体指标的估计(亦称参数估计)；②假设检验：如t检验、u检验、方差分析、χ^2检验、秩和检验等；③二项分布和Poisson分布的应用；④直线相关回归、协方差分析等。

(3) 临床流行病学中常用的统计方法　包括定群研究和病例对照研究资料分析中常用的统计方法；随访资料的生存率分析；诊断试验和判断一致性的统计方法；序贯试验等。

(4) 多因素分析的统计分析方法　包括多元回归和逐步回归；判别分析和聚类分析；logistic回归；Cox风险比例回归等等。

3．学习医学统计学应注意的问题

(1) 重点在于理解各种统计方法的基本概念，掌握适用范围和注意事项；学习过程中必须注意联系实际，结合专业。如联系医学文献和医学科研工作，评价其统计设计和分析的优缺点等。对于书中所引用的统计公式，只要求了解其意义及使用方法，不必深究其数理推导。

(2) 培养科学的统计思维方法。例如变异的客观存在，抽样误差的不可避免，混杂因素对试验结果的干扰，如何运用良好的设计避免偏倚，控制误差，假设检验的基本思想，根据概率作出统计结论的思想等。

(3) 本教材供基础、临床、口腔医学类专业的学生学完卫生学（或预防医学）中的医学统计方法之后，开设医学统计学选修课使用，目的在于提高学生医学统计水平，为今后深入进行科学研究打下良好的基础。因此，对于一些基本的统计方法，如平均数、率和比的应用、t检验、χ^2检验等，为了帮助读者更好地掌握内容，书中作了概略的回顾。本书还介绍了一些较为复杂的内容，如方差分析（二）、多变量分析等，目的是便于读者自学及便于今后工作需要时作参考。

4．统计工作的步骤和资料的类型

(1) 统计工作的进行可分为四步：设计、搜集、整理和分析资料。四个步骤互相联系，缺一不可。其中设计是整个统计研究的基础，在设计时应当对后三个步骤进行周密的考虑，并且在整个研究中自始至终地认真贯彻执行。

(2) 医学统计资料一般分为定量资料和定性资料(包括按等级分组的资料)，研究者必须根据不同资料类型选用适当的统计方法。

①定量资料 (quantitative data)　又称计量资料。对每一观察单位用定量的方法测定某项指标所得的资料称为定量资料。如调查某地成年男子（20—45岁）的平均血压，每个成年男性的血压测得值（mmHg，1mmHg=133.3Pa)，调查某地男性的红细胞平均数，每个成年男性红细胞的测得值（个/mm^3）等。

②定性资料 (qualitative data)　将观察单位按属性或类型分组计数所得的资料称为定性资料，根据属性或类型分组的多少又可分为下列两类：

a. 二项分类定性资料：将观察单位按两种属性分类，如死亡和生存；治愈和未愈；有效和无效等。

b. 多项分类定性资料：可分为两类。一类是将观察单位按多种属性分类，彼此之间互斥，如血型（A型、B型、AB型、O型等），这类资料应当选用定性资料的统计分析方法进行处理。另一类为各属性间有一种等级关系，如疗效观察可分为：治愈、显效，

好转、无效等，某些临床检验的结果分为－、±、＋、＋＋、＋＋＋、＋＋＋＋等的等级关系。此类按属性分组、各属性间又有程度（等级）差别的资料，称等级资料，应当选用适当的统计方法，如非参数统计方法（见第十二章12.1和12.2）进行处理。

研究者必须根据对观察单位测量的不同手段划分资料的类型，以便选择最佳统计分析方法，进行正确描述和推断。

5．几个基本概念

(1) 总体（population）与样本(sample) 总体是根据研究目的所确定的性质相同的所有个体的某种变量值的集合。例如调查某地成年男子的脉搏数，变量为脉搏数，每个个体所测得的脉搏数（次/分）为变量值，所有的脉搏数便组成了总体。这类有明确范围限制（如空间、时间）的总体称为有限总体；另一类不易划定确切范围的总体称为无限总体。如研究某药治疗糖尿病患者的效果，组成该总体的个体为各个糖尿病患者，研究者所设想的是所有的糖尿病患者，并无空间和时间的限制，因而是无限的。由于医学研究中的总体大都是无限总体，所以人们只能从中抽取一部分进行研究，并用研究的结果去推断总体。这些从总体中随机地抽取、进行研究的部分个体所组成的集合，称为样本。

(2) 参数（parameter）和统计量(statistics) 统计学中把总体的指标统称为参数。如研究某地成年男子的平均脉搏数（次/分），并从该地抽取1000名成年男子进行测量，所得的样本平均数即称为统计量。习惯上用希腊字母表示总体参数，例如μ表示总体均数，π表示总体率，σ表示总体标准差等。以拉丁字母表示统计量，$\bar{X}$表示样本均数，p表示样本率，s表示样本标准差等。

(3) 抽样误差（sampling error） 由于总体中的个体间往往存在着变异，随机抽取的样本仅是总体中的一部分个体，因而样本测得的指标（统计量）往往与总体指标（参数）存在着差异，这种由于随机抽样所造成的样本的统计量与总体参数的差异，称为抽样误差。

(4) 概率（probability） 概率是描述事件发生可能性大小的一个度量。例如事件A发生的可能性大小，称为事件A的概率，常记为$P(A)$，简称为P。它的取值范围由0到1。

1.2 平均数和标准差的应用

1．平均数的计算与应用 平均数主要用于反映一组观察值的平均水平，描述计量资料的集中趋势，亦称中心位置指标。在应用中应当根据分布的特点，选择适当的平均数。利用频数分布表识别分布的特点是一种较好的方法，例如表1.1就是描述某年某地102名7岁女孩身高（cm）的频数分布表，它由身高的组段和相对应的频数两列组成。下面介绍几种常用的平均数计算方法：

(1) 均数（average） 亦称算术平均数（arithmetic mean），简称均数（mean）。总体均数用希腊字母μ（读作*mu*）表示，样本均数用$\bar{X}$表示。它适用于对称性分布的资料，如例1.1（表1.1）中某市102名7岁女孩身高的频数分布，身高居中的组段频数较多，较高或较矮的频数逐渐减少，呈对称性分布。

计算式为：

表 1.1　某年某地102名7岁女孩身高(cm)的频数分布

身高(cm)	频数，f
100—	1
104—	4
108—	10
112—	20
116—	22
120—	20
124—	14
128—	6
132—	4
136—140	1
	102

① 不分组资料

$$\bar{X}=\frac{X_1+X_2+\cdots+X_n}{n}=\frac{\Sigma X}{n} \qquad (1.1)$$

式中$\bar{X}$为均数，X表示变量值，Σ为求和的符号。

② 分组资料　例数较多时进行分组的资料可用此法计算：

$$\bar{X}=\frac{\Sigma fX}{\Sigma f} \qquad (1.2)$$

式中f表示频数，X表示组中值，即每组的下限加上限除以2$\left(\frac{\text{下限}+\text{上限}}{2}\right)$，其它符号与式（1.1）相同。因此，该式计算的均数是一个近似值。

如表1.1中资料的均数为

$$\bar{X}=\frac{1\times102+4\times106+\cdots+4\times134+1\times138}{102}$$

$$=\frac{12160}{102}=119.22(\text{cm})$$

某市7岁女孩身高的均数为119.22cm。

(2) 几何均数（geometric mean）　几何均数适用于某些呈正偏态分布但经过数据转换为对数值后呈对称性分布和正态分布的资料。如表1.2中200名正常成人血铅含量的频数分布，可见资料呈正偏态分布；经对数转换后资料呈近似正态分布（见表1.3）。几何均数还可用于呈等比级数分组的资料，如医学实践中抗体的平均滴度和平均效价等。

几何均数的计算式分为：

① 不分组资料

$$G=\lg^{-1}\left(\frac{\lg X_1+\lg X_2+\cdots+\lg X_n}{n}\right)=\lg^{-1}\left(\frac{\Sigma\lg X}{n}\right) \qquad (1.3)$$

表 1.2 某年某地200名正常成人血铅含量的分布

血铅含量 (μg/100g)	频数
3—	47
9—	50
15—	44
21—	27
27—	18
33—	5
39—	5
45—	2
51—	1
57—63	1
	200

表 1.3 某年某地200名正常成人血铅含量经对数变换后的分布及均数计算表

$\lg X$ (1)	血铅含量，X (2)=$\lg^{-1}$(1)	人数，f (3)	组中值，$\lg X$ (4)	$f\cdot\lg X$ (5)=(3)(4)
0.45—	2.8—	1	0.525	0.525
0.60—	4.0—	15	0.675	10.125
0.75—	5.6—	20	0.825	16.500
0.90—	7.9—	32	0.975	31.200
1.05—	11.2—	36	1.125	40.500
1.20—	15.8—	48	1.275	61.200
1.35—	22.4—	28	1.425	39.900
1.50—	31.6—	16	1.575	25.200
1.65—1.80	44.7—63.1	4	1.725	6.900
合计	—	200	—	232.050

式中lg是以10为底的对数。

或 $$G=\sqrt[n]{X_1\cdot X_2\cdot X_3\cdots X_{n-1}\cdot X_n}$$

例1.2 用玫瑰花结形成试验，检查11名流行性出血热患者的抗体滴度，得资料如下：

1:20, 1:20, 1:80, 1:80, 1:320, 1:320, 1:320, 1:320, 1:320, 1:80, 1:80, 求患者抗体的平均滴度。

$$\lg G=\frac{5\times\lg 320+4\times\lg 80+2\times\lg 20}{11}=2.0673$$

$$G=\lg^{-1}2.0673=116.76$$

平均滴度为1:116.76。

② 分组资料 在例数较多或相同的观察值个数较多时宜用加权法计算，计算式为：

$$G=\lg^{-1}\left(\frac{\Sigma f\lg X}{\Sigma f}\right) \tag{1.4}$$

例1.3 （表1.3）某地200名正常人血铅含量的几何均数为

$$G=\lg^{-1}\left(\frac{232.050}{200}\right)=14.46(\mu g/100ml)$$

按公式（1.4）计算得200名正常人血铅含量的几何均数为14.46μg/100ml。

(3) 百分位数（percentile）和中位数（median） 百分位数是一种位置指标，以P_x表示。一个百分位数将一个数列（总体或样本）的观察值分为两部分，即有$x\%$的观察值比它小，有$(100-x)\%$的观察值比它大，所以百分位数是一个界值。一组数从小到大有序排列后被P_1，P_2，P_3，…，P_{99}等界值分为100等份，各含1%的观察值，从而也是一个数列的百等份的分割值。

中位数亦称位置平均数，是百分位数中的一个特例，即第50%位数。它将数列平分为两半，在正态分布的情况下，中位数应与均数相等。理论上，中位数可用于任何分布的定量资料，但日常工作中多用于描述不对称分布的集中趋势。特别是当分布末端无确定数据，如<5或>100等。当不能计算均数和几何均数时，中位数更显出它的优越性。由于中位数没有应用每个变量值的信息，只与居中的数值有关，因而波动较大且不够敏感。例1.5（表1.4）中可见205例伤寒患者潜伏期（天），分布不对称，宜选用中位数描述其集中趋势。

① 直接由原始数据计算中位数 首先将观察值按从小到大的顺序排列，然后按下式计算。

n为奇数时 $$M=X_{\left(\frac{n+1}{2}\right)} \tag{1.5}$$

n为偶数时 $$M=\left[X_{\left(\frac{n}{2}\right)}+X_{\left(\frac{n}{2}+1\right)}\right]\Big/2 \tag{1.6}$$

式中n为样本含量，$\left(\frac{n+1}{2}\right)$、$\left(\frac{n}{2}\right)$及$\left(\frac{n}{2}+1\right)$为该组有序数值中，观察值的位次，

$X_{\left(\frac{n+1}{2}\right)}$、$X_{\left(\frac{n}{2}\right)}$、$X_{\left(\frac{n}{2}+1\right)}$为相应位次上的观察值。

例1.4 10名1岁儿童的体重（kg）测得值为9.2，9.4，9.8，10.2，10.6，11.1，11.4，11.6，11.7，11.9，试计算其中位数。

$n=10$，为偶数，按式（1.6），

$$M=\left[X_{\left(\frac{10}{2}\right)}+X_{\left(\frac{10}{2}+1\right)}\right]\Big/2=(X_5+X_6)/2=(10.6+11.1)/2$$
$$=10.85\ (kg)$$

② 用频数表计算百分位数和中位数百分位数的计算式为

$$P_x=L+\frac{i}{f_x}\left(n\cdot x\%-\Sigma f_L\right) \tag{1.7}$$

式中f_x为P_x所在组段的频数，i为所在组段的组距，L为该组的下限，Σf_L为小于L各组段的累计频数。中位数即第50百分位数，故$x=50$，$M=P_{50}$。故式（1.7）可改写为

$$M=L+\frac{i}{f_M}\left(\frac{n}{2}-\Sigma f_L\right) \tag{1.8}$$

表 1.4 某年某地 205 例伤寒患者的潜伏期(天)

潜伏期(天) (1)	人数，f (2)	累计频数 (3)	累计频率(%) (4)
2—	26	26	12.7
4—	29	55	26.8
6—	42	97	47.3
8—	50	147	71.7
10—	48	195	95.1
12—	4	199	97.1
14—	2	201	98.0
16—	2	203	99.0
18—	1	204	99.5
20—22	$\frac{1}{205}$	205	100.0

例1.5 现用表1.4的资料计算中位数，P_{25}，P_{75}，$P_{2.5}$，$P_{97.5}$等百分位数。

根据公式（1.7）得：

$$P_{50}=8+\frac{2}{50}(205\times50\%-97)=8.22(\text{天})$$

$$P_{25}=4+\frac{2}{29}(205\times25\%-26)=5.74\ (\text{天})$$

$$P_{75}=10+\frac{2}{48}(205\times75\%-147)=10.28\ (\text{天})$$

$$P_{2.5}=2+\frac{2}{26}(205\times2.5\%-0)=2.39\ (\text{天})$$

$$P_{97.5}=14+\frac{2}{2}(205\times97.5\%-199)=14.87(\text{天})$$

P_{25}和P_{75}又称四分位数，P_{25}称下四分位数，以Q_L表示，P_{75}称上四分位数，以Q_U表示，它们之间的距离，称四分位数间距。通常也用四分位间距来描述资料的变异程度，此间距越大说明变异度越大。

研究者也常用$P_{2.5}$和$P_{97.5}$，P_5和P_{95}等估计频数的分布范围，医学科研工作中常用此区间估计偏态资料的分布范围，但例数不宜太少。

2．标准差的计算和应用 一组资料除描述集中趋势以外，还应说明其离散程度，只有二者结合起来，才能全面了解资料的分布情况。最常用的描述离散程度（亦称变异程度）的指标有以下几种，其中标准差应用最广泛。

(1) 方差（variance） 计算式为

$$\sigma^2=\frac{\Sigma(X-\mu)^2}{N} \tag{1.9}$$

式中σ^2为方差，它反映了每一个观察值与总体均数离差的平方的平均值，可以较全面地

反映该组资料的变异情况，由于计量资料都有测量单位（如kg，cm等），所以方差的单位是原始数据单位的平方。将方差σ^2开方即得标准差，它与原始数据单位相同。

(2) 标准差 (standard deviation)

$$\sigma=\sqrt{\frac{\Sigma(X-\mu)^2}{N}} \tag{1.10}$$

因为在抽样研究中总体均数μ是未知的，故常用样本均数$\bar{X}$估计，又由于$\bar{X}$常不等于μ，使得样本计算的方差平均偏小，故用$n-1$代替式中的N，所以样本标准差s为：

$$s=\sqrt{\frac{\Sigma(X-\bar{X})^2}{n-1}} \tag{1.11}$$

式中$n-1$称为自由度 (degree of freedom)，分子中的离均差平方和可改写为下式：

$$SS=l_{xx}=\Sigma(X-\bar{X})^2=\Sigma X^2-\frac{(\Sigma X)^2}{n}$$

计算标准差常用下列公式

不分组资料 $$s=\sqrt{\frac{\Sigma X^2-(\Sigma X)^2/n}{n-1}} \tag{1.12}$$

分组资料 $$s=\sqrt{\frac{\Sigma fx^2-(\Sigma fx)^2/\Sigma f}{\Sigma f-1}} \tag{1.13}$$

例1.6 测得12名口腔鳞状上皮癌患者头发中的锌的含量（μg/g），得资料如下，求其标准差。

142.0, 257.6, 393.8, 119.8, 271.2, 293.7, 169.7, 417.6, 249.4, 185.9, 325.6, 210.3。

$$s=\sqrt{\frac{\Sigma X^2-(\Sigma X)^2/n}{n-1}}=\sqrt{\frac{865949.84-(3036.6)^2/12}{12-1}}$$

$$=94.17\ (\mu g/g)$$

故计算得口腔上皮癌患者头发中锌含量的标准差$s=94.17$ (μg/g)。

例1.1中102例7岁女童身高的标准差

$$s=\sqrt{\frac{1454808-(12160)^2/102}{102-1}}=7.14\ (cm)$$

故计算得102例7岁女童身高的标准差$s=7.14$cm。

(3) 标准差的应用 标准差是重要的变异度指标，常用于：

① 说明变量值分布的离散程度，在两组均数相近，度量单位相同时，标准差越大，说明变量值的变异程度越大。

② 结合均数描述服从正态分布资料的分布特征。

③ 结合均数计算变异系数。

④ 结合样本含量计算标准误。

(4) 变异系数 (coefficient of variation) 变异系数用CV表示，即标准差s与均数$\bar{X}$之比用百分数表示，计算式为：

$$CV=\frac{s}{\bar{X}}\times 100\% \tag{1.14}$$

由于CV是一个百分比，没有测量单位，因而具有便于比较分析的优点，常用于：

① 比较均数相差悬殊的各组资料的变异度。如同一成人血压资料中收缩压和舒张压的变异程度。

② 比较度量衡单位不同的各组资料的变异程度。如同一年龄组儿童的身高（cm）和体重（kg）的变异度等。

3．正态分布的特征及应用 正态分布（normal distribution）是统计描述和推断中常用的一种理论分布，它有以下特征：

(1) 正态分布曲线是以均数为中心，左右完全对称的钟形曲线，在横轴均数所在处上方曲线位置最高（图1.1）。

(2) 正态分布有两个参数，即均数μ和标准差σ。μ为位置参数，σ是变异度参数。当σ恒定时，μ越大，曲线越向横轴右方移动；μ越小，则越向左方移动。当测量单位一致时，σ越大，表示数据越分散，曲线越低平，跨度越大；σ越小，表示数据越集中，曲线越高峭，跨度相对越小。因而知道μ和σ后，正态曲线就固定下来了。

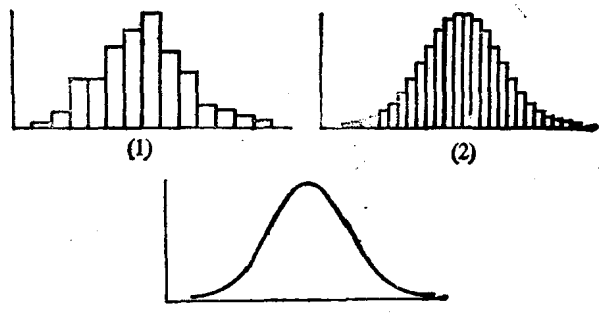

图 1.1 正态分布示意

(3) 为了便于描述和应用，常将正态分布的变量作数据转换。设$u=\frac{X-\mu}{\sigma}$使u的均数等于0，标准差等于1，即将μ的位置移到原点。横轴尺度以σ为单位，这样将正态分布变换为标准正态分布，u值又称为标准正态变量或标准正态离差（standard normal deviate）。

(4) 正态曲线下的面积有一定的分布规律（图1.2）。

理论上$\mu-\sigma$，$\mu+\sigma$占曲线下面积68.27%

$\mu-1.96\sigma$，$\mu+1.96\sigma$占曲线下面积95.00%

$\mu-2.58\sigma$，$\mu+2.58\sigma$占曲线下面积99.00%

统计学家已将标准正态分布曲线下的面积编制成工具表，在已知u值的情况下，可查得正态分布曲线下的面积，即概率（见附表1标准正态分布曲线下的面积），如已知$u=-0.82$，由表中可查得概率$P=0.2061$，由附表1右上角图形可见，此为正态曲线下单侧概率。当然若已知概率P，也可从表中倒查到u值。

不少医学测量值服从正态分布或近似正态分布，如同一年龄性别的身高，同性别的健康成人的红细胞等，均可按正态分布的规律估计正常值范围。表1.5为常用的u值表。

例1.7 某地调查360名成年男子的平均血红蛋白含量$\bar{X}=13.45$g/100ml，标准差$s=0.71$g/100ml，估计该地95%成年男子的血红蛋白的分布范围。

下限为 $\bar{X}-1.96s=13.45-1.96(0.71)=12.06$(g/100ml)

上限为 $\bar{X}+1.96s=13.45+1.96(0.71)=14.84$(g/100ml)

医务工作者可用此界值作为判断该地成年男子血红蛋白正常与否的参考值。

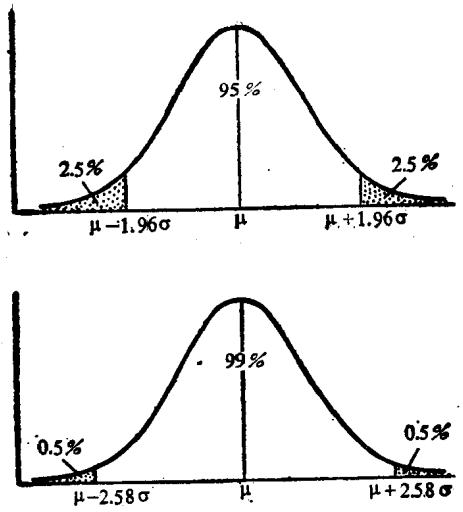

图 1.2 正态曲线下面积分布示意

表 1.5 常用的 u 值表（绝对值）

变量值分布（%）	单侧	双侧
80	0.842	1.282
90	1.282	1.645
95	1.645	1.960
99	2.326	2.576

虽然医学科学中不少的生理数值呈偏态分布，但其中大部分通过数据转换后可以呈正态或近似正态分布，因此，也可以用正态分布的规律来确定正常值范围。

1.3 总体均数的估计和假设检验

1. 均数的抽样误差—标准误 用样本的信息推断总体的特征，称为统计推断(statistical inference)。由于存在着抽样误差，样本统计量（$\bar{X}$）往往不等于总体参数（μ），因此，研究者应当了解抽样误差的大小。统计学中用标准误(standard error)对此进行衡量。

标准误可用下式计算

$$\sigma_{\bar{x}}=\frac{\sigma}{\sqrt{n}} \tag{1.15}$$

式中可见标准误的大小与总体标准差σ成正比，与观察例数的平方根成反比。在实际抽样研究中，往往不知道总体标准差σ，而是用单一样本的标准差 s 来估计 σ，于是将式(1.15)改写为

$$s_{\bar{x}}=\frac{s}{\sqrt{n}} \tag{1.16}$$