医用统计学基础

Foundation of Medical Statistics

主编　袁兆康　主审　刘汉强

江西高校出版社

书　名：医用统计学基础
主　编：袁兆康
出　版
发　行：江西高校出版社（南昌市洪都北大道 96 号）
经　销：各地新华书店
印　刷：南昌市印刷五厂
开　本：787×1092　1/16
印　张：9.25
字　数：250 千字
印　数：3000 册
版　次：1996 年 10 月第 1 版第 1 次印刷
定　价：11.00 元

ISBN7—81033—637—1
G·173

邮政编码：330046　电话：(0791)8513257 8512093 8519894

编 写 说 明

医用统计学对于提高医学生的素质，特别是科研能力所起的重要作用，已越来越为人们所重视。有些医学院校已为非预防医学专业医学生单独开设了医用统计方法这门课程，但课时较少。本书针对此特点并本着少而精的原则，重点介绍了医用统计学的基本原理和方法，使之适合于非预防医学专业医学生使用，使医学生能在较短的时间内，掌握医用统计学的基础知识，看懂有关杂志上的医学论文并能自己动手进行适宜的科研设计和数据处理。本书尽量从临床实例入手，阐述有关统计学的原理、概念、方法及注意事项，内容精炼、实用，语言通俗易懂，对非预防医学专业的医学生（包括临床、口腔、儿科、影像、护理、检验、基础医学等专业）来说，是一本较理想的医用统计学入门教材，同时也适用于广大医务工作者进行继续教育或业余自学。

本书共分11章，即绪论、计量资料的统计描述、计量资料的统计推断、计数资料的统计描述、计数资料的统计推断、非参数检验、直线相关与回归、半数效量、统计表与统计图、临床试验设计的指导原则、病例随访分析。在内容编排上，除考虑基本统计知识外，注意了临床统计的特点，增加了一些临床常用的统计方法。其中带有"*"的章节，主要目的在于增加本书的实用性，丰富和扩展读者的视野，只作为自学内容，课程上不作要求。

本书曾在江西医学院非预防医学专业学生中试用了两届，经广泛征求意见，又有江西中医学院的同行加盟，将原书进一步修改、补充、完善，现正式出版与读者见面。由于我们水平有限，加之时间仓促，书中难免会有不妥之处，敬望广大读者不吝赐教。

本书在编写、出版过程中，得到程本芳、邱逸樵、张令达、叶继红、余启胜等同志的大力支持和热情帮助，在此深表谢意。

袁兆康

1996年7月于江西医学院

目 录

第一章 绪 论

统计是国家实行科学决策和科学管理的一项重要基础工作，是认识国情国力、制定计划的重要依据。卫生统计学是把数理统计理论、方法应用于居民健康状况研究、医疗卫生实践和医学科学研究的一门应用学科，它侧重研究数据的搜集、整理和分析。

卫生统计学主要包括三方面的内容：

1. 医用统计方法及其基本原理；

2. 居民健康状况统计：包括医学人口统计、疾病统计和生长发育统计等；

3. 卫生服务统计：包括卫生资源、医疗卫生服务需求和利用、医疗保健制度等。

本教材重点介绍第一部分内容，使医学生掌握基本的统计理论与方法，以便在工作岗位上更好地总结卫生工作经验，更有效地从事医学科研工作及阅读医学文献。

第一节 统计工作的基本步骤

统计工作包括以下四个基本步骤：统计设计、统计资料的搜集、整理和分析。这四个步骤是紧密联系而不可分割的，任何一步发生差错，都会影响统计分析的质量。

一、统计设计(statistical design)

“凡事预则立，不预则废”，说明做任何工作，计划是很重要的。一个周密完善的统计设计，可以用较少的人力、物力和时间取得较好的成果，达到“事半功倍”的效果，因此，必须十分重视统计设计在医学科研和医疗卫生实践中的作用。统计设计的内容包括资料搜集、整理和分析全过程的总设想和安排，详见第十章。

二、搜集资料(collection data)

按设计的要求，及时取得准确、完整的原始资料(raw data)，是保证统计分析结论正确的关键一步。医学统计资料的来源主要有三个方面：

(一)统计报表：这是由国家统一制定，要求各个医疗卫生机构定期逐级填报的统计资料，如卫生基本情况年报表、医院工作年报表、居民病伤死亡原因年报表、传染病年(月)报表等。填写报表要做到及时、准确、完整，不得伪造统计数字，不得任意修改项目。

(二)经常性工作记录：如门诊病历、住院病历、健康检查记录、传染病报告卡、肿瘤报告卡、出生报告卡、死亡报告卡等。

应用医院病历时，应注意以下几个问题：

1. 医院病历具有挑选性，不能反映一般人群的特征。如大医院的住院病人往往病情较重或疑难病例较多等。

2. 一个地方医院病历资料的病例总数，不等于该地总的发病人数。所以，用医院病例数估计发病率时要注意这点。

3. 不同时期，不同医院病历记录的格式、繁简不一，所以在对比分析资料时应予以注意。

(三)专题调查或实(试)验：根据研究工作的需要组织专题调查或实验，可以在短期内获得有关的信息，但事先必须制定周密的调查(或实验)方案。

三、整理资料(sorting data)

在整理资料之前，应先对原始资料作详细、认真的检查核对，以保证原始资料的正确性和完整性。

原始资料记录发生错误的原因主要有二点：一是偶然填写错误；二是由于对填写的项目理解不准确引起的差错。后者的影响是大的。

检查核对资料有无错误，一般用两种方法：一是逻辑检查，即检查有无互相矛盾的地方，如性别填写"女"，死亡原因填写"前列腺癌"；年龄填写"2"，身高填写"165 厘米"；结婚次数填写"0"，生育胎次填写"2"等。二是计算检查，如纵横合计相加是否等于总计。如果把要检查的内容都编成计算机程序，计算机能按照指令对资料进行检查，但计算机一般只能检查出逻辑性错误。

整理资料的过程可归纳如下：

原始资料的检查审核 {人工审核 / 计算机审核} → 设计分组 {质量分组 / 数量分组 / 等级分组} → 拟整理表 → 归纳 {人工汇总 {划记法 / 分卡法} / 计算机汇总} → 统计表。

四、分析资料(analysis data)

资料经过整理后，便可按设计的要求和资料的性质，计算统计指标，必要时可做统计推断，并用适当的统计图表展示资料，以便阐明事物的内在联系和规律，最后结合专业作出恰如其分的解释和结论。详见有关章节。

第二节　统计资料的分类

统计资料按其性质可分为计数资料和计量资料两大类，介于其中的还有等级资料。不同类型的资料应采用不同的统计方法进行分析。

一、计数资料(enumeration data)

计数资料，又称属性资料(attribute data)，是将每个观察单位按其性质分类，然后清点各种性质观察单位数的资料。如用某药治疗一批痢疾病人，最后按治愈与未愈进行分组并清点各有多少病人；某人群中 O、A、B、AB 各种血型的人数等，均为计数资料。

二、计量资料(measurement data)

测量每个观察单位某项指标量的大小，所得的资料称计量资料，一般有度量衡单位。计量资料又可分为两类，一类是离散型或间断型，如口腔中龋齿的个数，妇女生育小孩的个数等，这种计数只能是 0 和正整数，不会是负值，也没有小数点；另一类是连续型，理论上在任何两个数值之间还有无穷多个数据，例如血清总胆固醇值 3.0～3.5mmol/L 之间理论上有无穷多个数据。

三、等级资料(ranked data)

各类之间有程度的差别，给人以"半定量"的概念，如疗效中的痊愈、显效、有效、无效；症状中的重、中、轻、无；化验中的一、十、卅、卌。这种资料其等级的排列是有序的。

根据分析的需要，上述三种资料可以互相转化，例如血红蛋白原属计量资料，若按血红蛋白正常与异常分组，得到各组人数，则为计数资料；若将血红蛋白按量的多少分为五级：

血红蛋白量(g/dL)	等　　级
∠6	重度贫血
6～	中度贫血
9～	轻度贫血
12～	血红蛋白正常
16～	血红蛋白增高

得到各等级人数，则为等级资料。有时亦可将计数资料或等级资料量化，使其转换成计量资料，如将男和女分别取 0 和 1；或将上述血红蛋白量的五个等级分别取 1、2、3、4、5。这时计数资料或等级资料就转化为计量资料。

第三节　统计中的几个基本概念

一、变量与变异

变量(Variable)是被观察单位(或个体)的特征，如身高、血压、体温、脉博等；性别相同、年龄相同的人，其身高、血压、脉博等都会有所不同，这种个体间的差别，通称为变异(Variation)。统计研究的对象是有变异的东西，如果研究的对象各个个体都完全相同，没有变异，也就没有必要进行统计研究了。

二、总体与样本

总体(population, universe)是根据研究目的确定的同质的研究对象的全体，更确切地说，是性质相同的所有观察单位某种变量值的集合。总体可分为有限总体和无限总体，例如研究某市 1995 年正常成人的血压，该市 1995 年全部正常成人便构成一个总体，这个总体包括的个体数是有限的所以是有限总体；有时总体是假想的，抽象的，例如研究某药治疗高血压患者的疗效，总体观察单位(个体)数显然是不确定的，故称无限总体。

样本(Sample)是从总体中随机抽取的部分观察单位某种变量值的集合。抽样研究的目的是由样本信息推论总体，所以，要求样本必须可靠和有代表性。

三、参数与统计量

根据总体观察单位求出的统计指标，称为参数(parameter)；根据样本观察单位求出的统计指标，称为统计量(statistic)。前者用希腊字母表示，后者用拉丁字母表示。

四、误差

误差(error)是实际测定值与真实值(理论值)之间的差值。医学研究中主要的误差有两种，即系统误差(systematic error)和随机误差(random error)。前者影响研究结果的准确度，后者影响研究结果的精密度，系统误差常以绝对误差、相对误差或回收率等指标来表示；随机误差通常用极差、标准差或变异系数等指标表示。

五、准确度和精密度

准确度(accuracy)和精密度(precision)是评价某种检测方法(包括仪器、试剂、操作方法等)可信赖程度的两项指标。准确度是指测定值与真值的接近程度，用以说明测定方法有无系统误差，一般用回收率表示，公式为：

回收率(%)　(检测值 - 本底值)/加入量 × 100%

回收率愈接近100%，则准确度愈高，当回收率偏离100%较大时，表示检测方法存在系统误差。精密度是指对同一标本进行多次重复测定时，测定值彼此接近的程度。一般用标准差或变异系数表示。标准差愈小，表示检测方法的重现性好，随机误差小。

在评价检测方法时，首先要考虑准确度，准确度差的方法，纵使精密度高，也是不能应用的。同样，方法的精密度差时，其准确度也不会高的。为了更直观理解准确度和精密度，现举例说明如下：

三人射击，以靶心为真值，三人各射击四发子弹，结果如下：

A的四发子弹的弹孔在靶上密集，可是，都靠在靶心的一侧较远处，其精密度高，但不准确；

B的四发子弹的弹孔分散在靶的上下左右，其精密度和准确度均差；

C的四发子弹的弹孔全部紧靠靶心，其精密度和准确度均好。

六、概率

统计理论中，概率(probability，P)的概念应用极广，任何统计分析结果均用一个 P 值表示。P 的概念是什么？按概率的统计定义，某事件的概率即是：在重复无数次的条件下，该事件的发生率。实际上所得到的概率只不过是这个理论概率的近似值，只是一个估计值。

P 值介于0～1之间，P 愈接近1，事件发生的可能性愈大；P 愈接近0，则发生的可能性愈小。在医学科学研究中，P 用得很多，例如，$t=2.9$，$P<0.02$，这里的 P 究竟代表什么？它代表机遇概率(chance probability)，或代表"检验假设"成立的概率。

在医学科研中，$P=1$ 的事件(称为必然事件)和 $P=0$ 的事件(称为不可能事件)是极少见的，大多数的事件，其 P 值介于0～1之间，即在一定条件下可能发生也可能不发生，这种事件称为随机事件或偶然事件。

小　结

医学统计是统计学的一个分支学科，它应用统计学的理论与方法，研究医学领域中数据搜集、整理和分析。

统计工作可分为统计设计、资料搜集、资料整理和资料分析四个步骤，各步之间紧密相联，任何一步发生差错，都将影响到最后的统计结论。

统计资料按其性质可分为计数资料、计量资料和等级资料。根据研究分析的需要，上述三类资料可以互相转换。

总体是同质研究对象的全体；样本是从总体中随机抽取的部分个体某种变量值的集合。反映总体特征的统计指标称参数；反映样本特征的统计指标称统计量。系统误差主要影响研究结果的准确度，随机误差主要影响研究结果的精密度。评价检测方法时，应首先考虑准确度。概率是在大量重复的条件下，某事物的发生率，其值介于0～1之间。

(刘汉强)

第二章　计量资料的统计描述

用恰当的统计指标描述所获资料的数量特征，称统计描述(statistical discription)。计量资料的统计描述分两方面，一方面描述变量值向某一点集中的趋势，称集中趋势(central tendency)；另一方面描述变量值之间不相同的趋势，称离散趋势(tendency of dispersion)。

第一节　集中趋势的测量

描述计量资料集中趋势的统计指标称平均数(average)。它是统计中应用最广泛，最重要的一个指标体系，常用于描述一组同质计量资料的集中趋势，反映一组变量值的平均水平。常用的平均数有算术平均数、几何平均数、中位数等。

一、算术平均数(arithmetic mean)

算术平均数简称均数(mean)，是日常工作中用得最多的一种平均数，适用于对称分布的资料。总体均数用希腊字母 μ 表示，样本均数用 $\bar{x}$ 表示。

(一)直接计算法

当变量值(数据)个数不多时，可将各变量值相加，除以变量值的个数，即得均数。计算公式如下：

$$\bar{x} = \frac{x_1 + x_2 + \cdots + x_n}{n} = \frac{\sum x}{n} \tag{2.1}$$

式中 $\bar{x}$ 表示均数；Σ(读作 sigma)是总和的符号；x_1、x_2、…、x_n 表示各个变量值，n 表示变量值的个数，即样本含量。

例 2.1　测得 5 个周岁儿童的头围(cm)如下：44、45、46、47、48，求其平均头围。

将数据代入式(2.1)：

$$\bar{x} = \frac{44 + 45 + 46 + 47 + 48}{5} = \frac{230}{5} = 46(\text{cm})$$

(二)频数表计算法

当变量值个数较多时，用直接法计算较繁，且容易出错，这时可先编出频数表(frequency table)，再用加权法计算均数。

例 2.2　某地 130 名正常成年男子红细胞(万/mm^3)如表 2.1，求其均数。

计算步骤如下：

1. 编制频数表

(1)求全距(range，简写为 R)：　$R = x_{max} - x_{min}$　　(2.2)

式中 R 表示全距；x_{max} 表示最大值；x_{min} 表示最小值。

本例：$R = 588 - 379 = 209$(万/mm^3)

(2)确定组段数(k)

一般分 8～15 组段为宜。组数太多，显示不出资料的规律性，且计算较繁；组数太少，误差较大。本例定 $k = 10$

表 2.1　某地区 130 名正常成年男子红细胞数(万/mm³)

379 *	457	519	486	428	467	537	498	445	588 * *
453	516	484	415	466	531	497	443	477	478
510	483	411	463	528	494	440	474	567	505
481	398	461	523	490	435	470	546	503	440
389	457	521	487	420	467	538	498	446	478
454	516	485	417	466	532	497	443	477	507
513	483	413	464	529	495	442	474	569	453
418	401	462	526	491	436	473	549	504	478
394	457	523	490	431	468	539	499	448	508
454	517	486	427	466	536	498	443	477	453
515	484	413	464	529	496	442	475	569	480
482	410	462	526	493	439	474	561	504	510
398	458	523	490	433	468	540	500	449	480

* 最小值　* * 最大值

(3)求组距(i)：　$i = R/k$　　(2.3)

本例：$i = \frac{209}{10} = 20.9$(万/mm³)

因组数是人为确定的，本身容许有一定程度的波动，因此组距也容许出现一定误差，本例为计算方便，定组距为 20 万/mm³。

(4)划分组段

每一个组都应当有一个起始值作为组下限(low limit)和一个终止值作为组上限(up limit)。第一组段应包括最小值，最后一个组段应包括最大值。因上一组段的上限就是下一组段的下限，为避免两组段界限互相包含，组段常用各组段的下限及短线"~"表示。各组段的上限都归到了下一组段，最后一组段应同时写出其下限和上限。组段划分的情况见表 2.2 第(1)栏。

表 2.2　某地区 130 名正常成年男子红细胞数频数分布及均数计算表

红细胞(万/mm³)	划　记	频　数(f)	组中值(x)	fx
(1)	(2)	(3)	(4)	(5)
370~	下	2	380	760
390~	正	4	400	1 600
410~	正下	9	420	3 780
430~	正正正一	16	440	7 040
450~	正正正正下	22	460	10 120
470~	正正正正正	25	480	12 000
490~	正正正正一	21	500	10 500
510~	正正正下	17	520	8 840
530~	正下	9	540	4 860
550~	下	4	560	2 240
570~590	一	1	580	580
合计	—	130	—	62 320

(5)划记时表

通过划"正"字，将各变量值划归各组，见表 2.2 第(2)栏。清点"正"字即得到分布于各组段的变量值个数，称之为频数，用"f"表示，见表 2.2 第(3)栏。

2. 计算均数

从表 2.2 可看到，130 个变量值分布于 11 个组段，虽然我们知道每一个组段有几个变量值，但仅从频数表中我们并不知道每个变量值的具体数值。可以设想，如果我们获得的样本是随机的，那么变量值在每个组段中的分布也应是随机的，没有趋向性。因此我们可取每个组段中间的那个值作为分布于这个组段中的所有变量值的均数，称作这一组段的组中值(class mid - value)用 x 表示。计算公式为：

$$x = \frac{\text{本组段下限} + \text{本组段上限}}{2} \tag{2.4}$$

各组段组中值的计算结果见表 2.2 第(4)栏。用各组的组中值与该组的频数相乘，即可得到分布于该组的变量值的合计数，即 fx，见表 2.2 第(5)栏。将各组的合计数相加，即可得到所有变量值的总合计数，即 Σfx。用 Σfx 除以总样本含量 Σf，即可得到所要求的均数。计算公式为：

$$\bar{x} = \frac{f_1x_1 + f_2x_2 + \cdots + f_kx_k}{f_1 + f_2 + \cdots + f_k} = \frac{\Sigma fx}{\Sigma f} \tag{2.5}$$

将表 2.2 计算结果代入式(2.5)：

$$\bar{x} = \frac{62320}{130} = 479.38(\text{万}/\text{mm}^3)$$

这里各组的频数起着"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，这种计算方法称为加权法(weighting method)

二、几何平均数(geometric mean)

当一组计量数据的各变量值之间是倍数关系时(即呈等比级数变化)，应按倍数平均，用几何平均数表示其平均水平。几何平均数用 G 表示，适用于对数对称分布的资料，常用来计算抗体的平均滴度，急性传染病的平均潜伏期，某事物平均发展速度等。

(一)直接计算法

几何平均数是几个变量值的连乘积开 n 次方所得的根。计算公式为：

$$G = \sqrt[n]{x_1 \cdot x_2 \cdots x_n}$$

两边取对数得：

$$\lg G = \frac{\lg x_1 + \lg x_2 + \cdots + \lg x_n}{n} = \frac{\Sigma \lg x}{n}$$

因此：

$$G = \lg^{-1}\left(\frac{\Sigma \lg x}{n}\right) \tag{2.6}$$

例 2.3　有 5 人的血清效价分别为 1∶10，1∶100，1∶1 000，1∶10 000，1∶100 000，求其平均效价。

将数据代入式(2.6)：

$$G = \lg^{-1}\left(\frac{\lg 10 + \lg 100 + \lg 1\,000 + \lg 10\,000 + \lg 100\,000}{5}\right)$$

$$= \lg^{-1}\left(\frac{1 + 2 + 3 + 4 + 5}{5}\right)$$

$$= \lg^{-1}3 = 1\,000$$

即其血清抗体平均效价为1:1 000。

(二)频数表计算法

当数据较多时,可像计算算术平均数那样,先编频数表,再用加权法计算几何均数。计算公式为:

$$G = \lg^{-1}\left\{\frac{\Sigma f \lg x}{\Sigma f}\right\} \tag{2.7}$$

例2.4 40名麻疹易感儿接种麻疹疫苗后一个月,血凝抑制抗体滴度见表2.3第(1)、(2)栏,求平均滴度。

表2.3 例2.4平均滴度的计算

抗体滴度	人数,f	滴度倒数,x	$\lg x$	$f\lg x$
(1)	(2)	(3)	(4)	(5)=(2)(4)
1:4	1	4	0.602 1	0.602 1
1:8	5	8	0.903 1	4.515 5
1:16	6	16	1.204 1	7.224 6
1:32	2	32	1.505 1	3.010 2
1:64	7	64	1.806 2	12.643 4
1:128	10	128	2.107 2	21.072 0
1:256	4	256	2.408 2	9.632 8
1:512	5	512	2.709 3	13.546 5
	40			72.247 1

先按表2.3第(1)、(2)栏的数据求其x、$\lg x$及$f\lg x$、$\Sigma f\lg x$见表2.3第(2)~(5)栏,将计算结果代入式(2.7):

$$G = \lg^{-1}\left\{\frac{72.247\,1}{40}\right\} = \lg^{-1}1.806\,2 = 64$$

即平均滴度为1:64。

三、中位数与百分位数

把一组数据按大小顺序排列,位次居中的那个数据即这一组数据的中位数(median),用M表示。理论上,中位数可用于任何分布的资料,但日常工作中多用于描述分布不对称资料的集中趋势,特别是当分布的首尾无确定数据时(如<5或>100等),中位数更显示出它的优越性。由于中位数只与居中的数据有关,未应用每个数据的信息,因而波动较大且代表性较差。

(一)直接计算法

首先将变量值按大小顺序排列,然后按下式计算:

n为奇数时 $M = x_{(\frac{n+1}{2})}$ (2.8)

n为偶数时 $M = [x_{(\frac{n}{2})} + x_{(\frac{n}{2}+1)}]/2$ (2.9)

式中n为样本含量,$(\frac{n+1}{2})$,$(\frac{n}{2})$及$(\frac{n}{2}+1)$为该组有序数值中,变量值的位数。

例2.5 某病有患者7人,其潜伏期(天)分别为5,6,7,8,9,10,20,求其中位数。

本例样本含量为7，是奇数，将数据代入式(2.8)：

$$M = x_{(\frac{7+1}{2})} = x_4 = 8(\text{天})$$

例2.6 某病有患者8人，其潜伏期(天)分别为5，6，7，8，9，10，20，25，求其中位数。

本例样本含量为8，是偶数，将数据代入式(2.9)：

$$\begin{aligned} M &= [x_{(\frac{8}{2})} + x_{(\frac{8}{2}+1)}]/2 \\ &= (x_4 + x_5)/2 \\ &= \frac{8+9}{2} \\ &= 8.5(\text{天}) \end{aligned}$$

(二)频数表计算法

当数据较多时，也可先编频数表，再按频数表计算中位数。其公式为：

$$M = L + \frac{i}{f_m}(\frac{n}{2} - \Sigma f_L) \qquad (2.10)$$

式中 L 表示中位数所在组的下限；i 表示中位数所在组的组距；f_m 表示中位数所在组的频数；n 表示样本含量，亦即总频数 Σf；Σf_L 表示小于 L 的各组累计频数。

例2.7 某年某地205例伤寒患者的潜伏期如表2.4，求其平均潜伏期。

表2.4 某年某地205例伤寒患者的潜伏期(天)

潜伏期 (1)	人数，f (2)	累计频数 (3)
2～	26	26
4～	29	55
6～	42	97
8～	50	147
10～	48	195
12～	4	199
14～	2	201
16～	2	
18～	1	
20～22	1	
	205	

首先确定中位数所在组，由 $\frac{n}{2}$ 找出中间位次，按频数表计算累计频数到略大于 $\frac{n}{2}$ 为止，即中位数所在组。本例：$\frac{n}{2} = \frac{205}{2} = 102.5$，“6～”组累计频数为97，尚未超过102.5，“8～”组累计频数为147，大于102.5，故中位数所在组为“8～”组，将有关数据代入式(2.10)：

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

平均潜伏期为8.22天。

(三)百分位数(percentile)

把一组数据按大小顺序排列成一个数列，并将其分为100等份，每一等份各含1%的数

据，与第 x 等份相对应的值，称第 x 百分位数，用 P_x 表示。实际上，中位数就是第 50 百分位数 P_{50}。因此，百分位数的计算与中位数类似，其公式为：

$$P_x = L + \frac{i}{f_x}(n \cdot x\% - \Sigma f_L) \tag{2.11}$$

式中 P_x 表示第 x 百分位数，L 表示百分位数所在组的下限；i 表示百分位数所在组的组距；f_x 表示百分位数所在组的频数；n 表示样本含量；Σf_L 表示小于 L 的各组累计频数。

仍以例 2.7 为例，求其 P_{25}，P_{75}，$P_{2.5}$，$P_{97.5}$。

首先确定百分位数所在组，由 $n \cdot x\%$ 求出百分位数在具体数据中的位次，按频数表累计频数至略大于 $n \cdot x\%$ 为止，即百分位数所在组。

本例：$205 \times 25\% = 51.25$

$205 \times 75\% = 153.75$

$205 \times 25\% = 5.13$

$205 \times 97.5\% = 199.88$

故其百分位数所在组分别为“4～”组，“10～”组，“2～”组和“14～”组，将有关数据代入式 (2.11)：

$$P_{25} = 4 + \frac{2}{29}(205 \times 25\% - 26) = 5.47(\text{天})$$

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

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

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

百分位数是一种位置指标，主要用以表示某一特定位置上具体数值的大小，以此来划定变量值的频数分布范围。例如 P_{25} 表示全部变量值中，有 25% 的变量值小于此水平，75% 的变量值大于此水平；同理，P_{75} 表示全部变量值中，有 75% 的变量值小于此水平，25% 的变量值大于此水平。照此类推，$P_{2.5}$～$P_{97.5}$ 之间应包含全部变量值的 95%，医学科研中常用此区间来估计偏态分布资料的频数分布范围，有些正常值范围（指大多数正常人某个变量值的分布范围）就是用百分位数来确定的。例如 95% 的正常值范围，如过大过小均异常，中间正常，其范围为 $P_{2.5}$～$P_{97.5}$；如仅过大异常，则超过 P_{95} 为异常；仅过小异常，则低于 P_5 为异常。

第二节　离散趋势的测量

一组资料除描述集中趋势外，还应说明其离散程度，只有将二者结合起来，才能全面了解资料的分布情况。

例 2.8　有三组数据如下：

Ⅰ：4, 5, 6, 7, 8

Ⅱ：2, 3, 6, 9, 10

Ⅲ：4, 5, 6, 6, 6, 6, 7, 8

这三组数据的算术平均数都是 6，但其数量特征并不相同，因此有必要描述其离散趋势。常用的描述离散趋势的指标有：全距，四分位数间距，方差和标准差。

一、描述离散趋势的常用指标

(一)全距(R)　又称极差,是一组变量值中最大值与最小值之差,其计算见式(2.2)。以例2.8三组数据为例,全距分别为4,8,4。第Ⅰ组和Ⅲ组数量特征并不相同,但全距未反映出它们的差别。可见用全距描述数据的离散趋势虽然简单明了,计算方便,但它仅考虑了资料的最大值与最小值,没有考虑其它变量值,因此不够全面且不稳定,易受两端数据的影响。

(二)四分位数间距　将一列数据由小到大按顺序排好,再分成四等分,每一等分包含1/4的数据,第一个四分之一等分处的数值称下四分位数,用Q_L表示;第三个四分之一等分处的数值称上四分位数,用Q_u表示。Q_u-Q_L即四分位数间距。与全距相比,四分位数间距剔除了两头各1/4的数据,因此比较稳定。常用于描述一些非对称分布资料的离散趋势。即论上,$Q_u-Q_L=P_{75}-P_{25}$,故计算四分位数间距常用百分位数法。

(三)方差　比较理想的描述离散趋势的指标,应考虑到每个变量值的离散程度　能否用每一变量与均数之差的和即离均差总和$\Sigma(x-\bar{x})$来表示呢?对于对称分布的资料,由于正负相消,$\Sigma(x-\bar{x})=0$。如例2.8第Ⅰ组数据$\Sigma(x-\bar{x})=(4-6)+(5-6)+(6-6)+(7-6)+(8-6)=-2-1+0+1+2=0$。为了克服这一缺点,可考虑把每个$(x-\bar{x})$平方后再相加,即$\Sigma(x-\bar{x})^2$,这叫离均差平方和(sum of squares)。仍以例2.8第Ⅰ组的数据为例,$\Sigma(x-\bar{x})^2=(4-6)^2+(5-6)^2+(6-6)^2+(7-6)^2+(8-6)^2=(-2)^2+(-1)^2+0^2+1^2+2^2=10$。用同样的方法得到第Ⅱ组、第Ⅲ组数据的离均差平方和分别为50和10

离均差平方和的大小除了与数据离散趋势的大小有关外,还与样本含量有关,样本含量越大,变量值个数越多,离均差平方和必然越大。因此就出现了诸如例2.8第Ⅰ组和第Ⅲ组数据尽管数量特征不同,但离均差平方和相同都是10的结果。为了便于在变量值个数不相等时离散趋势的比较,可将离均差平方和除以变量值个数,所得结果称方差(variance)。

总体方差用σ^2表示,计算公式为:

$$\sigma^2=\frac{\Sigma(x-\mu)^2}{N} \tag{2.12}$$

样本方差用S^2表示,根据数理统计研究结果,用样本算得的方差往往比总体方差偏小,为了得到总体方差的较好估计值,计算样本方差时常将分母中的n减去1,故:

$$S^2=\frac{\Sigma(x-\bar{x})^2}{n-1} \tag{2.13}$$

$n-1$在统计上称为自由度(degree of freedom),用ν表示。自由度是统计上的常用术语,其意义是随机变量能"自由"取值的个数。如有一样本$n=4$,$\bar{x}=5$,在自由确定4,2,5三个数据后,第四个数据只能是9,否则$\bar{x}\neq 5$。因此在这里自由度$\nu=n-1=4-1=3$。

(四)标准差　变量值是有单位的,为了用同样的单位表示其离散趋势,将方差开平方,就得到了标准差(standard deviation)。标准差是一种比较理想的,最常见的表示离散趋势的指标。

总体标准差用σ表示,计算公式为:

$$\sigma=\sqrt{\frac{\Sigma(x-\mu)^2}{N}} \tag{2.14}$$

样本标准差用S表示,计算公式为:

$$S=\sqrt{\frac{\Sigma(x-\bar{x})^2}{n-1}} \tag{2.15}$$

公式(2.15)中的离均差平方和常用 SS 或 l_{xx} 表示，数学上可以证明：

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

由此，式(2.15)可改写为：

$$S = \sqrt{\frac{\Sigma x^2 - \frac{(\Sigma x)^2}{n}}{n-1}} \tag{2.16}$$

式(2.16)是日常工作中常用的计算公式。

二、标准差的计算

(一)直接计算法

例2.9 仍以例2.1 5个周岁儿童的头围为例，列计算表2.5。

表2.5 5名周岁儿童头围的标准差计算表

x	x^2
44	1936
45	2025
46	2116
47	2209
48	2304
$\Sigma x = 230$	$\Sigma x^2 = 10590$

将表2.5计算结果代入式(2.16)：

$$S = \sqrt{\frac{10590 - \frac{230^2}{5}}{5-1}} = 1.58(\text{cm})$$

(二)频数表计算法

加权法计算标准差的公式为：

$$S = \sqrt{\frac{\Sigma fx^2 - \frac{(\Sigma fx)^2}{\Sigma f}}{\Sigma f - 1}} \tag{2.17}$$

例2.10 仍以例2.2 130名正常成年男子红细胞数为例，列计算表2.6。

表2.6 130名正常成年男子红细胞数标准差计算表

红细胞数($万/mm^3$)	x	f	fx	fx^2
370～	380	2	760	288 800
390～	400	4	1 600	640 000
410～	420	9	3 780	1 587 600
430～	440	16	7 040	3 097 600
450～	460	22	10 120	4 655 200
470～	480	25	12 000	5 760 000
490～	500	21	10 500	5 250 000
510～	520	17	8 840	4 596 800
530～	540	9	4 860	2 624 400
550～	560	4	2 240	1 254 400
570～590	580	1	580	336 400
合　计		130	62 320	30 091 200