全国高等医药院校试用教材

（供卫生专业用）

卫生统计学

四川医学

全国高等医药院校试用教材

（供卫生专业用）

卫 生 统 计 学

主编单位

四 川 医 学 院

编写单位

山 西 医 学 院　　上海第一医学院

北 京 医 学 院　　哈尔滨医科大学

武 汉 医 学 院　　湖 南 医 学 院

四 川 医 学 院

人 民 卫 生 出 版 社

卫生统计学

四川医学院　主编

人民卫生出版社出版

人民卫生出版社印刷厂印刷

新华书店北京发行所发行

787×1092毫米16开本　15¾印张　366千字

1978年12月第1版第1次印刷

1981年12月第1版第4次印刷

印数：57,151—76,950

统一书号：14048·3661　定价：1.25元

编写说明

本书是卫生部组织编写的高等医学院校《卫生统计学》教材，供卫生专业试用。

全书共十七章：第一章绪论；第二章至第十四章讲述基本的统计方法。其中，第二至第十二章先讲统计分析方法，并专章讨论了计算半数致死量和确定正常值范围问题。在此基础上，第十三章和第十四章再讲述调查和实验的全面设计，这样便于学生理解，并牢固树立收集资料、整理资料和分析资料全过程的整体概念；第十五章至第十七章讲述应用上述统计方法研究人民生老病死的健康统计问题。书中记有"*"号部分可供学生选读和自学参考。书末有四个附录。其中："附表"是统计分析计算常用的工具表；"习题"可供各校安排课间实习选题的参考；"英汉名词对照"，供学生阅读英文书刊的参考；为了便于学生能在中学数学基础上，了解若干统计公式的数学原理，列入了"部分统计公式的推导"，供自学参考。

为了集思广益，在审稿、定稿时，我们邀请了部分在教学、科研和卫生防疫工作中有经验的同志参加讨论。其中有第四军医大学郭祖超教授，军事医学科学院薛仲三教授，中国医学科学院卫生研究所田凤调同志，中山医学院黄小兰、王志瑾同志，山东医学院王均乐同志，重庆医学院张照寰、周燕荣同志等。他们对本书的编写热情支持，提出了宝贵意见。此外，周燕荣、黄光明、康晓平、董景五、李艳华、杨珉等同志为本书的绘图、缮写和校对等工作，付出了辛勤的劳动，谨此致谢。

本书是解放后卫生部统一组织编写的第一本《卫生统计学》教材，限于我们的水平和缺乏编写经验，一定有不少缺点和错误，热忱欢迎读者批评指正，并希各校在使用过程中，不断总结经验，提出宝贵意见，以便进一步修改提高。

目　录

第一章 绪 论

§1.1 卫生统计学的任务与内容

辩证唯物主义认为：世界是物质的，物质是运动的，运动是有规律的，规律是可知的。为了能动地改造世界，首先要认识世界，研究物质世界运动的客观规律，这就要认真进行调查研究。毛泽东同志教导我们："没有调查就没有发言权"，又教导我们要"胸中有'数'。这是说，对情况和问题一定要注意到它们的数量方面，要有基本的数量的分析。任何质量都表现为一定的数量，没有数量也就没有质量。"毛泽东同志关于调查研究、数量与质量的辩证关系的教导，对卫生工作和医学科学研究工作均具有普遍的指导意义。

卫生统计学是在马列主义、毛泽东思想的指导下，把统计理论和方法应用于医疗卫生实践和医学科学研究的一门学科。如应用于研究社会条件、环境因素及生物因素对人民健康的影响，应用于评价医疗卫生措施的质量和效果，应用于医学科研的调查、实验设计和数据处理。并在广泛应用的基础上，不断发展统计理论与方法。电子计算机的普及，对卫生统计的发展提供了广阔的前景。电子计算机使各种信息的贮存和分析自动化，为繁复的，大量的资料收集、保存、整理和分析工作提供了极为有利的条件，这将把卫生统计工作提高到新的水平。卫生统计学范围较广，但本课程教学内容，仅限于收集、整理和分析资料的基本统计理论与方法，以及评价人民健康状况的疾病统计和死亡统计等。

卫生医师从事卫生防疫工作，必须根据人群健康与疾病状况采取防治措施，并评价其效果，还必须大力开展医学科学研究工作。卫生统计学是认识人群的健康与疾病现象数量特征的重要工具，是进行医药卫生科学研究的重要手段。因此卫生医师学习卫生统计学是十分必要的。

学习卫生统计学，要着重理解卫生统计学的基本原理和基本概念，掌握收集、整理与分析资料的基本知识与技能。重视原始资料的完整性与可靠性。对数据的整理与分析必须持严肃、认真和实事求是的科学态度。应用卫生统计方法，要结合专业知识，不断提高分析问题与解决问题的能力。对于本书中的数理统计公式，不必深究其数学原理，只要求弄清公式的适用条件和用法，并对计算结果能作出正确的分析。

§1.2 卫生统计工作的步骤与统计资料的类型

一、卫生统计工作的步骤

收集资料、整理资料和分析资料是卫生统计工作的三个基本步骤。

收集资料就是根据研究目的，进行调查或实验设计，然后按设计要求，收集准确与完整的原始资料。这是统计工作的前提与基础。

整理资料就是把收集到的原始资料，有目的、有计划地进行科学的加工，使资料系统化、条理化，以便进行统计分析。

分析资料就是把统计整理的结果，进一步计算相应的指标，结合专业知识，运用统计方法进行分析对比，阐明事物的规律，应用于改造世界的实践。

必须指出，这三个基本步骤是密切联系不可分割的，任何一个环节有了缺陷，都会影响研究结果的正确性。

二、统计资料的类型

卫生统计资料一般分为计量资料与计数资料两大类。不同的统计资料应采用不同的统计分析方法。

计量资料是对每个观察单位用定量方法测定某项指标的数值大小所得的资料，一般用度量衡等单位表示，如身高(cm)、体重(kg)、浓度(mg/l)、脉搏(次/分)、血压(mmHg)、白细胞总数(个/立方毫米)。这类资料可用第二、三、四、八、九等章的分析方法。

计数资料是先将观察单位按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。例如对某小学校全体学生进行粪检蛔虫卵，将粪检结果分为阳性与阴性两组，然后清点得阳性组人数与阴性组人数。又如调查某人群的血型分布，按照A、B、AB、O四型分组，清点得该人群各血型组的人数。这类资料可用第五、六、七等章的分析方法。

在医学实践中，有些资料具有计数资料的特性，但同时又兼有半定量的性质，我们称它为按等级分组资料。按等级分组资料是将观察单位按某项指标的等级顺序分组，再清点各组观察单位的个数所得的资料。如检查一组急性病毒性肝炎病人血清的麝香草酚絮状试验，将检查结果按－、＋、++、+++、++++等级分组，清点得每组病人数。又如观察某病的治疗效果，将患者按疗效的等级分为痊愈组、显效组、无效组、恶化组、死亡组，再清点得各组的人数。这类资料可用第九章等的统计分析方法。

根据分析研究的目的，有时可以把计量资料变换为计数资料或按等级分组资料。例如白细胞总数属于计量资料，但可以按白细胞总数正常(4000～10000个/立方毫米)与白细胞总数不正常(多于10000个/立方毫米或少于4000个/立方毫米)分为两组，清点各组人数，就成为计数资料；如果按白细胞总数过高(多于10000个/立方毫米)、正常(4000～10000个/立方毫米)、低下(少于4000个/立方毫米)分为三组，清点各组人数，这样就成为按等级分组的资料了。

§ 1.3 几个基本概念

一、总体与样本

根据研究目的确定的研究对象的全体称为总体，样本是总体中有代表性的一部分。例如要调查某年某地区12岁健康男孩的身高水平，那么该地区全部12岁健康男孩就

是一个总体。我们从中随机抽取120名进行身高测量，这120名12岁健康男孩就称为样本。通过计算这120名12岁健康男孩的平均身高，就可以运用统计方法估计该地区全部12岁健康男孩的身高水平。这种从总体中随机抽样，用样本指标估计总体指标的方法叫抽样方法。在抽样过程中为了避免主观意愿或客观无意识的偏性影响，使样本能够充分反映总体的情况，必须遵循"随机化"的原则，使每个研究对象被抽取的机会完全均等。抽样方法在卫生防疫工作及医学科学研究中被广泛地应用。

二、误　差

统计上所说的误差，泛指测得值与真值之差，以及样本指标与总体指标之差，主要的有下列三种。

1．系统误差　在收集资料过程中，由于仪器不准，标准试剂未经校正，医生掌握疗效标准偏高或偏低等原因，可使观察结果成倾向性的偏大或偏小，这叫系统误差。系统误差影响原始资料的准确性，应力求避免。如已发生，要尽力查明其原因，予以校正。

2．随机测量误差　在收集资料过程中，即或方法统一，仪器及标准试剂已经校正，但由于各种偶然因素的影响造成同一对象多次测定的结果不完全一致，这种误差往往没有固定的倾向，而是有的稍高有的稍低，叫随机测量误差。随机测量误差是不可避免的，但应努力作到仪器性能及操作方法稳定，使其控制在一定的允许范围内。必要时可作统计处理。

3．抽样误差　即使消除了系统误差，并把随机测量误差控制在允许范围内，样本平均数(或率)与总体平均数(或率)间仍可能有差异，这种差异叫抽样误差，这是由于个体差异造成的。如居住在同一地区的12岁健康男孩，他们的身高总是有高有矮，这些个体差异是客观存在，不可避免的。因此，我们从该地区12岁健康男孩中随机抽取一个120人的样本，如算得他们的平均身高为143.1 cm。这个样本指标不一定正好等于该地所有12岁健康男孩的真实平均身高(总体指标)，这叫抽样误差。抽样误差有一定的规律，研究和运用抽样误差的规律，进行调查(或实验)设计与资料分析，是卫生统计学的重要内容之一。

三、概　率

概率是指某事件出现的可能性。这种可能性的大小在数学上用分数、小数或百分数来表示。例如某地区统计某年出生数88273名，其中女婴占49.05%，男婴占50.95%。通过上述大量统计观察结果，女婴或男婴与出生总数之比约为常数0.5，这时我们就说生男或生女的概率约为0.5。但是该地区某产院每天出生男婴或女婴与出生总数之比不一定都是0.5，有时甚至偏离0.5较远，这是由于观察数量较少，偶然性所致。

统计上用符号"P"来表示概率，概率P的数值波动介于0与1之间。某一事件必然不发生，则该事件发生的概率为0；某一事件必然发生，则该事件发生的概率为1。某事件发生的概率愈接近0，表示该事件发生的可能性愈小；概率愈接近1，表示事件发生的可能性愈大。统计分析的许多结论，都是建立在概率大小的基础上的。

（凌瑞珠、林琼芳　编）

第二章　平均数与标准差

§ 2.1　平　均　数

平均数是分析计量资料常用的一种统计指标。它说明一组观察值的平均水平或集中趋势。

在卫生防疫工作和医学科研中常用的平均数有均数(又称算术平均数)、几何均数和中位数等。

一、均　　数

当所处理资料的观察值大小分布比较对称时，可计算均数。均数常用$\bar{X}$表示。

(一) 直接计算法　当观察值的个数不多时，可直接将各观察值相加，除以观察值的个数，算出均数。

例 2.1　有 8 名正常人的血清球蛋白含量(g%)分别为 3.1，2.5，2.9，2.6，2.6，3.0，2.6，2.6，求这 8 名正常人的平均球蛋白含量。

$$\bar{X}=\frac{3.1+2.5+2.9+2.6+2.6+3.0+2.6+2.6}{8}=\frac{21.9}{8}=2.74(g\%)$$

用公式表示：

$$\bar{X}=\frac{X_1+X_2+X_3+\cdots+X_n}{n}=\frac{\sum_{i=1}^{n}X_i}{n}=\frac{\Sigma X}{n} \qquad (2.1)$$

$$i=1,2,3,\cdots,n$$

式中$\bar{X}$为均数，$X_1, X_2, X_3, \cdots, X_n$为各观察值，$\Sigma$(读作 Sigma)为总加的符号，$n$为观察值的个数。

(二) 用频数表计算法　当观察值个数较多时，用直接计算法较繁，容易出错。这时可先编出频数表，再用加权法或简捷法计算均数。

例 2.2　1973年某市 120 名 12 岁男孩身高(cm)测量资料如下，求其均数。

142.3	156.6	142.7	145.7	138.2	141.6	142.5	130.5	132.1	135.5
134.5	148.8	134.4	148.8	137.9	151.3	140.8	149.8	143.6	149.0
145.2	141.8	146.8	135.1	150.3	133.1	142.7	143.9	142.4	139.6
151.1	144.0	145.4	146.2	143.3	156.3	141.9	140.7	145.9	144.4
141.2	141.5	148.8	140.1	150.6	139.5	146.4	143.8	150.0	142.1
143.5	139.2	144.7	139.3	141.9	147.8	140.5	138.9	148.9	142.4
134.7	147.3	138.1	140.2	137.4	145.1	145.8	147.9	146.7	143.4
150.8	144.5	137.1	147.1	142.9	134.9	143.6	142.3	143.3	140.2
125.9	132.7	152.9	147.9	141.8	141.4	140.9	141.4	146.7	138.7
160.9	154.2	137.9	139.9	149.7	147.5	136.9	148.1	144.0	137.4
134.7	138.5	138.9	137.7	138.5	139.6	143.5	142.9	146.5	145.4
129.4	142.5	141.2	148.9	154.0	147.7	152.3	146.6	139.2	139.9

1．编制频数表

（1）找出观察值中的最大值、最小值和全距。本例最大值为160.9 cm，最小值为125.9 cm，最大值与最小值之差称为全距，即160.9 cm－125.9 cm＝35 cm。

（2）根据全距决定组段数、组段和组距。一般分8～15个组段为宜。组段数过多，计算较繁，组段数过少则误差较大。本例若分10个组段，则组距为3.5 cm$\left(\frac{\text{全距}}{\text{组段数}}=\right.$组距，本例，35 cm÷10＝3.5 cm$\left.\right)$，为简便计，组距取整数为4 cm。第一组段要包括最小的观察值，最后一组段要包括最大的观察值。所以第一组段可从125开始，组距为4，即125-，129-，133-，···157-161（此组段包括了最大值160.9）。“125-”这个组段，表示120名男孩中凡是身高在125 cm至未满129 cm的人均应归入该组段。125 cm为本组段的下限，129 cm为本组段近似的上限。

（3）列表划记。划分组段后，记入表2.1第(1)栏，经划线记数得第(3)栏的频数分布。表2.1称为频数分布表，简称频数表。

2．计算均数

（1）用加权法计算均数。从表2.1看出身高在“125-”组段内有1人，在“129-”组段内有4人···等等。同一组段内每个人的身高是不相等的，我们可取组中值代表该组段每个人的身高。组中值$=\frac{\text{本组段下限}+\text{下组段下限}}{2}$。所以第一组段的组中值为$\frac{125+129}{2}$＝127，第二组段的组中值为$\frac{129+133}{2}$＝131，余类推，见表2.2第(2)栏。各组段内第(2)栏组中值X与第(3)栏频数f的乘积为第(4)栏的fX，将各组段的fX相加得ΣfX。再将此值除以总频数即得120名男孩的平均身高。因为这里的频数起到了“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。所以这种计算均数的方法称为加权法。

表2.1　1973年某市120名12岁男孩身高的频数分布表

组段 (1)	划记 (2)	频数 (3)
125-	一	1
129-	丅	4
133-	正丅	9
137-	正正正正正下	28
141-	正正正正正正正	35
145-	正正正正正丅	27
149-	正正一	11
153-	丅	4
157-161	一	1
		120

表2.2　120名12岁男孩身高均数的加权法计算表

组段 (1)	组中值X (2)	频数f (3)	fX (4)＝(2)(3)
125-	127	1	127
129-	131	4	524
133-	135	9	1215
137-	139	28	3892
141-	143	35	5005
145-	147	27	3969
149-	151	11	1661
153-	155	4	620
157-161	159	1	159
		120(Σf)	17172(ΣfX)

用公式表示：

$$\bar{X}=\frac{f_1X_1+f_2X_2+f_3X_3+\cdots+f_kX_k}{f_1+f_2+f_3+\cdots+f_k}=\frac{\Sigma fX}{\Sigma f} \qquad (2.2)$$

式中 $\Sigma f=n$。$X_1, X_2, X_3, \cdots, X_k$，分别为第一组段至 k 组段的组中值；$f_1, f_2, f_3, \cdots, f_k$，分别为第一组段至 k 组段的频数；ΣfX 为各组段内组中值与频数乘积的总和。

本例 $\Sigma f=120$ $\Sigma fX=17172$

$$\bar{X}=\frac{\Sigma fX}{\Sigma f}=\frac{17172}{120}=143.10\ (\text{cm})$$

即 1973 年某市 120 名 12 岁男孩的平均身高为 143.10 cm。

(2) 用简捷法计算均数。简捷法计算均数是在加权法基础上进一步简化的一种方法。其算法如下：

1) 先列计算表(表 2.3)。表中第(1)、(2)、(3)栏同表 2.2。

表 2.3 均数的简捷法计算表

组 段	组中值X	频数 f	$x=\frac{X-X_0}{i}$	fx	
(1)	(2)	(3)	(4)	(5)=(3)×(4)	
125-	127	1	−4	−4	−62
129-	131	4	−3	−12	
133-	135	9	−2	−18	
137-	139	28	−1	−28	
141-	143	35	0	0	
145-	147	27	1	27	+65
149-	151	11	2	22	
153-	155	4	3	12	
157-161	159	1	4	4	
		120		+3	

2) 选"假定均数"(以 X_0 表示)。一般选频数较多，并且比较中间的组中值为假定均数，本例可选 143 为假定均数。

3) 计算缩减值 x。将各组的组中值减去假定均数 143，再除以组距 4，求得缩减值 x。见第(4)栏。如第一组段的 $x=\frac{127-143}{4}=-4$，最后一组段的 $x=\frac{159-143}{4}=4$，余类推。由于组距相等，x 值是很有规律的。假定均数所在组段的 $x=0$，组中值小于假定均数各组段的 x 值依次为 $-1, -2, -3, \cdots$，组中值大于假定均数各组段的 x 值依次为 $1, 2, 3, \cdots$，所以 x 值可直接写出，不必通过计算。x 值是简化后的组中值，故称缩减值。

4) 求 Σfx。各组段内第(3)栏 f 与第(4)栏 x 的乘积为第(5)栏的 fx，其合计值即 Σfx。

5) 代入式(2.3)求均数。

$$\bar{X}=X_0+\frac{\Sigma fx}{\Sigma f}(i) \qquad (2.3)$$

式中 X_0 为假定均数，f 为频数，x 为缩减值，i 为组距。

将表 2.3 有关数值代入式(2.3)得

$$\overline{X}=143+\frac{3}{120}\times 4=143.10\ (cm)$$

用简捷法算得的均数与加权法结果一致，而方法比较简便。

二、几何均数

几何均数又称几何平均数，常用G表示。医学和卫生学中计算抗体的平均滴度，抗体平均效价，食物中毒的平均潜伏期等，习惯上用几何均数。因为这些观察值往往彼此相差较大，有的资料甚至成倍数关系，用均数表示其平均水平时受少数特大或特小值影响较大。

例 2.3　有 5 人，其血清抗体效价分别为 1:10,1:100,1:1000,1:10000,1:100000。求其效价倒数的平均水平。

若计算其均数

$$\overline{X}=\frac{10+100+1000+10000+100000}{5}=22222$$

明显看出，由于受特大值的影响，均数偏在大值一边，所以用均数反映这类资料的平均水平是不合适的。

几何均数就是 n 个观察值的乘积开 n 次方所得的根。计算公式为：

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

$$\lg G=\frac{\lg X_1+\lg X_2+\lg X_3+\cdots+\lg X_n}{n}=\frac{\Sigma\lg X}{n}$$

$$G=\lg^{-1}\left(\frac{\Sigma\lg X}{n}\right)\qquad(2.4)$$

本例计算几何均数

$$\lg G=\frac{\lg 10+\lg 100+\lg 1000+\lg 10000+\lg 100000}{5}$$

$$=\frac{1+2+3+4+5}{5}=3$$

$$G=\lg^{-1}3=1000。$$

由此可见，本例用几何均数 1000 反映平均水平是较合适的。

(一) 直接计算法

例 2.4　某地 10 人接种某种疫苗后，测定抗体滴度如下：1:2,1:2,1:4,1:4,1:4,1:4,1:8,1:8,1:16, 1:32。求该疫苗的抗体平均滴度。

先求观察值倒数的几何均数，代入式(2.4)

$$G=\lg^{-1}\left(\frac{\lg 2+\lg 2+\lg 4+\lg 4+\lg 4+\lg 4+\lg 8+\lg 8+\lg 16+\lg 32}{10}\right)$$

$$=\lg^{-1}0.7526=5.7$$

该疫苗的抗体平均滴度为 1:5.7。

(二)用频数表计算法　当观察值个数较多时，可先编频数表，再按频数表计算几何

均数。

例 2.5 某菌苗接种后二周，受试者的血清凝集效价资料如表 2.4 第(1)、(2)栏所示，计算其平均凝集效价。

表 2.4 平均凝集效价的计算

凝集效价 X (1)	人数 f (2)	lgX (3)	$f$lgX (4)=(2)(3)
20	6	1.3010	7.8060
40	9	1.6021	14.4189
80	21	1.9031	39.9651
160	7	2.2041	15.4287
320	5	2.5051	12.5255
	48		90.1442

计算公式为

$$G=\lg^{-1}\left(\frac{\Sigma f\lg X}{\Sigma f}\right) \qquad (2.5)$$

将表 2.4 有关数值代入式(2.5)得

$$G=\lg^{-1}\left(\frac{90.1442}{48}\right)=\lg^{-1}1.8780=75.51$$

故平均凝集效价为 1:75.51。

三、中 位 数

把一组观察值按大小顺序排列，位次居中的那个数值即中位数。在它的上下各有相等的频数分布着。当一组观察值中，大部分较集中，只有少数的甚至个别的分散在一侧，或资料的分布情况不清楚，或数据一端（或两端）无界限时（如观察记录只有大于或小于多少，而无确切值），宜用中位数表示它们的集中趋势，如传染病的平均潜伏期等。中位数用 M 表示。

(一) 直接计算法

1. 观察值的个数(n)为奇数时。位次居中的那个观察值即中位数。

例 2.6 有 9 个某种传染病人，他们的潜伏期（天）分别为 2，5，4，3，3，6，9，16，3。求中位数。

先将数据按大小次序排列，得

2 3 3 3 4 5 6 9 16

本例，$n=9$ 是奇数，所以位次居中（即第五位）的观察值“4”就是中位数。

2. 观察值的个数(n)为偶数时。位次居中的两个观察值的平均数即中位数。

例 2.7 某医生随机抽取某工厂轧钢工人 20 名，测得其白细胞核棘突百分比如下，求中位数。

0 0 0 0 1 2 2 4 4 5

6 6 6 7 8 9 10 11 13 14

本例，$n=20$ 是偶数。所以位次居中（即第 10 位和第 11 位）的两观察值为 5 和 6，其平均数$\frac{5+6}{2}=5.5$即中位数。

(二) 用频数表计算 当观察值的个数较多时，可先将资料编成频数表，再用式(2.6)求中位数。

例 2.8 现有某钢铁工厂轧钢工人 204 人，分别测得其血中大单核细胞百分数，编成频数表如表 2.5 第(1)、(2) 栏资料，求其中位数。

按频数表计算累计频数至略大于$n/2$(即中间位次)为止，也就是中位数所在的组段。如本例应累计至略大于204/2=102为止。计算累计频数的方法见表2.5第(3)栏："0-"组段仍为24，"2-"组段为24+40=64，"4-"组段为64+55=119，已略大于102，不再累计。该组频数$f_M=55$，该组下限L=4，小于L的累计频数C=64，组距$i=2$。代入式(2.6)即求得中位数。

$$M=L+\frac{i}{f_M}\left(\frac{n}{2}-C\right) \qquad (2.6)$$

本例 $M=4+\frac{2}{55}\left(\frac{204}{2}-64\right)=5.38(\%)$

注意　式(2.6)实际是按比例内插的算式。

表 2.5　204名轧钢工人血中大单核细胞百分数的中位数计算表

分组 (1)	频数 (2)		累积频数 (3)
0-	24		24
2-	40		64
4-	55	38	119
		17	
6-	37		
8-	27		
10-	18		
12-	1		
14-	0		
16-	1		
18-	0		
20-	1		
	204		

最后应当强调指出，平均数的计算与应用必须建立在同质基础上。如研究儿童生长发育时，要在儿童同性别同年龄的基础上计算其平均身高，平均体重。研究某种职业毒害对工人健康的影响时，要按职业和工龄分组计算平均数和进行分析。研究某病的平均潜伏期，必须是确诊为该病的对象。正如马克思在资本论中指出的那样："平均量只是种类相同的许多不同的个别量的平均"。列宁也曾指出："将大小作坊混合在一起而得出的'平均'数字是完全荒谬的"。在医学科学研究中也必须遵循这个原则，如将不同质的事物混在一起计算平均数是毫无意义的，并且是有害的。

§2.2 标准差

一、标准差的意义

对一组观察值进行分析时，不仅要计算平均数，反映其平均水平，还要用一些指标反映其变异程度的大小。例如有两组数据：

甲组　98　99　100　101　102

乙组　80　90　100　110　120

两组的均数都是100，但分布的情况却不同。甲组比较集中，即变异较小，而乙组比较分散，即变异较大。所以对一组观察值的描述，除了说明其平均水平外，还要说明其变异程度的大小。表示平均水平的指标用平均数，前已述及。表示变异程度大小的指标用全距、方差、标准差。最常用的是标准差。

1．全距　又称极差。是一组观察值中最大值与最小值的差。如上述两组数值，甲组的全距=102-98=4，乙组的全距=120-80=40。甲组全距较小，乙组全距较大，说明甲组数据比较集中，乙组数据比较分散。

用全距表示变异程度的大小简单明了，但它仅考虑了资料的最大值与最小值，而没

有考虑其他数值，因此是不够全面的。

2．方差　要克服全距的缺点，必须全面考虑到每一个观察值。能否用每一观察值与均数之差的总和简称离均差总和，即 $\Sigma(X-\bar{X})$ 来表示呢？但由于正负相消，离均差总和等于 0，即 $\Sigma(X-\bar{X})=0$。如上例：

甲组　$(98-100)+(99-100)+(100-100)+(101-100)+(102-100)=0$

乙组　$(80-100)+(90-100)+(100-100)+(110-100)+(120-100)=0$

这样仍不能表示变异程度的大小。为了进一步克服这一缺点，可考虑把每个$(X-\bar{X})$平方后再相加，简称离均差平方和，即 $\Sigma(X-\bar{X})^2$，这样就避免了正负相消的问题。如

甲组　$$\begin{aligned}\Sigma(X-\bar{X})^2&=(98-100)^2+(99-100)^2+(100-100)^2+(101-100)^2\\&\quad+(102-100)^2\\&=10\end{aligned}$$

乙组　$$\begin{aligned}\Sigma(X-\bar{X})^2&=(80-100)^2+(90-100)^2+(100-100)^2+(110-100)^2\\&\quad+(120-100)^2\\&=1000\end{aligned}$$

但离均差平方和的大小除了与变异程度大小有关外，还与观察值的个数有关。观察值的个数越多，则 $\Sigma(X-\bar{X})^2$ 就越大，所以还应当取其均数，即方差，用 S^2 表示。

$$S^2=\frac{\Sigma(X-\bar{X})^2}{n}$$

但根据数理统计研究结果，用样本资料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，可将分母中的 n 减去 1，故

$$S^2=\frac{\Sigma(X-\bar{X})^2}{n-1} \qquad (2.7)$$

式中　$n-1$ 在统计上称为自由度。

由式(2.7)容易理解，方差愈小说明观察值的变异程度愈小，方差愈大，说明变异程度愈大。

3．标准差　方差是将每个离均差平方后总加起来被自由度除。因此，观察值的单位也都进行了平方，如身高原来的单位是 cm，而方差的单位就是 cm^2，为了用同样单位来表示，所以又把方差开平方，这就是标准差，以 S 表示。算式为：

$$S=\sqrt{\frac{\Sigma(X-\bar{X})^2}{n-1}} \qquad (2.8)$$

又可写为

$$S=\sqrt{\frac{\Sigma X^2-\frac{(\Sigma X)^2}{n}}{n-1}} \qquad (2.9)$$

用式(2.9)计算 S 不必先求 $\bar{X}$，较为简便。

和方差一样，标准差愈小，说明观察值的变异程度愈小，标准差愈大，说明变异程度愈大。