高等学校试用教材

应用统计学

华东化工学院　吴乙申　主编

机械工业出版社

高等学校试用教材

应用统计学

华东化工学院　吴乙申　主编

机械工业出版社

高等学校试用教材
应 用 统 计 学
华东化工学院　吴乙申　主编
*
机械工业出版社出版（北京阜成门外百万庄南里一号）
（北京市书刊出版业营业许可证出字第117号）
北京市密云县印刷厂印刷
新华书店北京发行所发行·新华书店经售
*
开本 787×1092 1/16·印张 20 1/2·字数 498 千字
1986年11月北京第一版·1986年11月北京第一次印刷
印数　0,001—7,450·定价 3.45 元
*
统一书号：15033·6550

前　言

应用统计是高等学校管理工程专业的基础课程之一，也是进行管理决策所需运用的重要数量分析工具。历次高等学校管理工程专业会议都曾讨论过应用统计学的教学大纲和内容，但均未成书。本书是根据1983年4月机械工业部管理工程专业教材编审委员会杭州会议上确定的教学大纲编写的。

本书由上海华东化工学院管理工程系吴乙申担任主编，上海复旦大学管理学院丁伯金和上海同济大学管理学院王永安担任协编。参加编写工作的有吴乙申（绪论及第一、二章）、丁伯金（第九、十、十一、十二章）、王永安（第三、四、五章）、上海交通大学管理学院盛宝忠（第八章）、栾军（第七章）、陈俊芳（第六章）。最后由主编作了统一的修改。

本书由武汉华中工学院林少宫担任主审。参加审稿工作的还有清华大学管理学院李端敏、华中工学院应用数学系樊孝述、南京华东工程学院管理工程系漳渭基和机械工业部教材编辑室责任编辑崔国徽同志。

全书共分十二章。主要内容包括：第一章统计描述，第二章抽样与抽样分布（这一章是由描述统计过渡到推断统计的桥梁），第三章参数的点估计，第四章参数的区间估计，第五章假设检验（这三章是统计推断的基本内容），第六章方差分析，第七章试验设计，第八章回归分析与相关分析（这三章是处理两个以及多个变量之间的关系的主要统计方法），第九章非参数方法（这一章讲述独立于统计分布即对总体不作任何假设的一些统计方法），第十章抽样检验方法（这一章主要讲产品质量检验中应用的一些抽样检验方法），第十一章经济指数，第十二章时间序列分析（这两章讲述社会经济统计中经常应用的统计分析方法）。

本书是高等学校管理工程类专业的试用教材。适当裁减一部分内容后可供管理干部专修科试用，同时也可作为高等学校应用数学专业联系实际的数理统计教材或补充读物。此外，还可作为在职统计人员、预测人员、决策人员、质量管理人员、科学试验人员掌握统计学知识的自学读物。学习本书所需要的预备知识只限于工科院校本科生学习的高等数学、概率论和线性代数。

在本书编写过程中曾蒙同济大学管理学院名誉院长翟立林和上海机械学院副院长戴鸣钟的重视关怀和大力支持，并承有关兄弟院校的多方协助。华东化工学院管理工程系的郁斐章及李静同志也花了不少劳动。特在此一并表示衷心的感谢。

由于编者水平所限，再加时间仓促，疏漏谬误之处在所难免，恳切希望广大读者给予指正。

编　者

1985年2月

目　录

绪　论

应用统计学是一门以随机现象为研究对象的学科。

运用抽样调查或科学实验所获得的数据，根据概率论的原理，进行分析论证，引出结论，作为解决具体问题的决策依据。这些就是应用统计学所应研究的主要内容。

随机现象广泛存在于各门学科中。无论是自然科学或社会科学，哪里出现不确定性（随机性）的问题，哪里就需要统计分析的方法。这样，应用统计就成了其它学科的研究工具，并且渗透到农业、生物、物理、化学、医学、教育、心理、经济、政治、人口、社会学、管理以及工程技术等各个学科领域。近几年来电子计算机软件工程的不断进展，又为各种统计方法的应用提供了先进的手段，于是统计方法的应用更为普及了。

对统计学的认识有两种偏见我们认为都是不正确的。一种偏向是认为统计学的主要任务是搜集和整理数据以及把这些数据用表格或图形表示出来，或许还包括计算平均数、总数、百分比之类。另一种偏向是忽视统计描述工作的重要性，认为统计应以统计推断为主，包括回归分析、相关分析等等所涉及的统计推断。实际上统计描述和统计推断是紧密相连的，是一个整体。例如，模拟随机抽样就需要经验分布的概念，而后者却正是统计描述的一个内容。所以，忽视其中任何一方都是不应该的。

本书名为《应用统计学》，那么，应用统计与数理统计有区别吗？我们认为两者的基本内容是一致的。区别只在于数理统计在阐述统计原理及方法时偏重于抽象的数学方法，而应用统计在介绍统计的原理和方法时，密切联系应用的具体问题。

我们希望通过这本应用统计学的学习，能使学生和广大管理人员正确掌握统计学的思想和方法，并在各自的专业领域中广泛应用统计学这个有力的工具。

以上就是我们编写本书的出发点和愿望。

第一章 统 计 描 述

§1-1 统 计 学 概 述

一、统计学和统计分析

统计的应用，早在封建时代就已开始。由于征收赋税、徭役和兵丁的需要，历代政府都很重视土地和人口的统计，并有详尽的记载。虽然统计的产生已有很长的历史，但它真正成长为一门成熟的学科并作为其它学科广泛使用的一种工具，不过是最近半个世纪以来的事情。今天，无论是对自然科学还是对社会科学的各个分枝学科，如生物、物理、化学、医药、气象、水文、经济、金融、财贸、管理和工程技术等等来说，统计学已经成为它们不可缺少的工具了。人们已经习惯于阅读统计数字，搜集统计资料，进行统计试验，用以作出有意义的分析，获得合理的推断，并作为制定最优决策、提高经济效益的依据。

统计学是一门应用性很强的学科。这可以通过下面的两个例子来说明。例如：假设某工厂要选举厂长，竞选者有两人，究竟谁能当选，须待选举结果方知分晓。如欲在事前摸摸情况，估计一下谁可能当选，可以采用民意测验的办法。具体做法是：任选一部分选民（其数目可为数十个或者上百个不等，一般视选民总人数的多寡而定），由他们进行预选，统计选举结果，得票百分率较高的便是获胜者。拿这个结果作为对未来选举结局的预测，一般是相当可靠的。又例如：某药厂发明了一种新针剂，欲知其预防某种疾病的疗效。于是选择了数目均等的两部分人，在一部分人身上注射新药液，另一部分人身上不注射此种药液。经过一段时间，统计这两部分人感染某种疾病的百分率，如果注射过新药的人感染率远较未经注射的人低，那末，就可认为这种新针剂对预防此种疾病有疗效。根据什么理由，可以得出这样的结论呢？以后的统计分析将回答这个问题。这两个例子仅说明统计学具有应用性很强的特征。以后随着本书内容的展开，将会碰到更多的统计学应用实例。

统计学是通过对数据进行搜集、整理、约简、分析和推断从而引出结论的一门学科。对数据进行加工的整个过程叫统计分析。同时，根据对数据加工的性质的不同，统计学又可分为描述性统计和推断性统计两个部分，统计分析的全过程包括了这两个部分。

二、描述性统计和推断性统计

在搜集数据之后，把它们整理成表格、图形的形式，并用一些特征值如平均数、百分数等来显示其面貌。这部分内容称为描述性统计。在统计学发展的初级阶段，统计学的内容大体上以描述性统计为主。它的作用是能将数据化为醒目的形式，快速地为我们提供有用的信息。这部分统计内容是统计学的先驱部分。由于我们采集的一组数据，一般只是总体数据中一个部分或者叫做样本，所以，怎样由样本所给的信息来窥测总体的面貌，就构成了统计学另一部分更重要的内容。这后一部分内容叫做推断性统计。由于统计学具有从部分推断总体的特征，就使统计学成为其它学科广泛应用的一种工具。

三、总体、个体和样本

我们在研究过程中获得的一组数据，如果是研究对象的全体所有数据，那末，这组数据

构成的集合就叫做总体，其中的每个元素叫做个体。由部分个体所组成的集合叫做样本，后者在一定的意义上“代表”了总体。例如，1982年6月30日，我国全国人口普查所获得的资料就构成研究中国人口问题的总体，其中每个人的资料，就是个体；同日江西省的人口资料，叫做样本。江西省的人口资料在一定意义上“代表”了中国人口的总体，所以我们称它为对应于全国人口资料的样本。

总体可以分成有限总体和无限总体两种，按其所包含个体的数目为有限和无限而区别。这样说并不意味着个体数目很大就是无限总体。社会学、经济学中有些研究资料的数目很大，如中国人口的数字接近十亿，但这个总体还是有限的。至于无限总体，往往是指一个连续过程所产生的结果充满在一定范围之内。例如，某刀具的切削长度可以在某个实数区间〔a，b〕内变化，这样形成的〔a，b〕实数区间，就包含着无穷多个实数。所以，我们称〔a，b〕为无限总体。

来自某个总体的样本，一般包含n个元素(个体)，称它为容量n的样本（或者称大小为n的样本）。通过样本所提供的信息以推测总体的容貌，这就是统计推断的主要内容。统计推断是以样本为基础的，并要运用概率原理进行论证。当然，我们也可以从总体直接引出结论。本章将首先介绍适用于总体和样本的两项统计描述。

四、统计分析过程框图

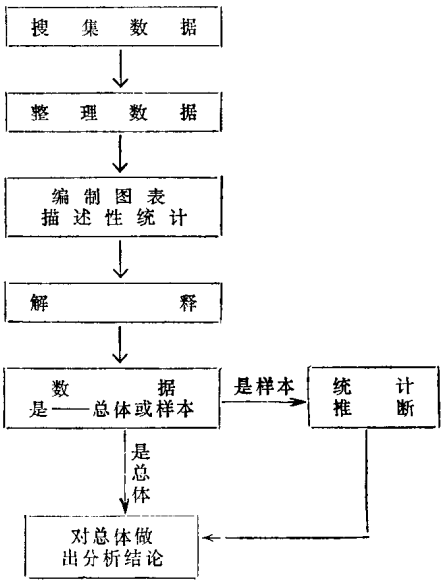

图1-1 统计分析过程示意图

按照上面对统计学的介绍，我们可用图1-1的框图来总结统计分析的过程。

§1-2 数据的搜集和整理

我们的出发点是假设研究工作者已经获得了一批统计资料，例如一个班级学生某门课程的考试成绩，某医院一组婴儿的体重记录，某橡胶厂所制轮胎寿命试验的记录。我们称这些统计资料为数据。作为一个统计工作者的职责首先是怎样把数据整理和组织起来，以为我们提供有用的信息。例如，把整理过的数据制作成看起来一目了然的图形或表格，从而得出对研究对象有意义的论断。

表示研究对象的变量可分为两大类型：一类是类别变量，另一类是数值变量。举例来说，如果我们要研究某地区人口的职业分布，则表示不同职业的变量就是类别变量。这种类别变量是按各人所属的职业表示的，如医生、教师、营业员、理发员等等。如果我们所研究的变量须用数值表示，例如人的体重、体温、高度等等，便称为数值变量。数值变量又可分为离散的和连续的，按照它们的取值是整数值或实数值来区分。在描述统计中，两大类型变量，都由频率分布图或其它条形图表示。在推断性统计中，所论述的主要变量将是数值变量。

§1-3 频率分布及其图形

一、频率分布

假设某商店在一个暑期内（共 40 天）每天销售的啤酒数如表 1-1 所列。

表1-1 某商店40天中日销售啤酒数记录

63	68	71	74	76	78	81	84	85	89
66	70	73	75	76	79	82	84	85	90
67	71	73	75	76	79	82	85	86	92
68	71	74	75	77	79	84	85	86	94

表 1-1 中的数据已经经过整理，由上而下、从左到右、由小到大排列着。现在我们要用频率表的形式把它们概括显示出来。方法是：首先把它们分组，然后列出落在每组内的频数，最后再计算频率，即相对频数，形成一个频率表。这里，首先要解决的一个问题，就是数据应分成几组。一般规则，组数应在 5 至 15 范围内。如果数据的总数少，组数也要相应少一些；如果数据的总数大，组数也应多一些。同时，每组的上下限应多设一位数，以便使记录的每一个数据都能明确应安放在哪一组内。例如在表 1-1 中最小的数据是 63，我们分组时应把组限扩展一位数，定为 59.5～64.5，(59.5 为下限，64.5 为上限)，使 63 能明确地落在这一组中。又如瓶数 70，根据表 1-2 的设组，明确地落在 69.5～74.5 这一组中。在表 1-2 中，我们把组数分成 7 组，每组组距为 5 。这样，组的全距为 35，近似于最大瓶数 94 与最小瓶数 63 之差。

表 1-2 是根据表 1-1 的数据组成的频率表。表中第一列表示组距；第二列为组中心；第三列为日销售瓶数落在对应组内的个数，即频数。例如，落在第一组 59.5～64.5 内的瓶数只有 63 一个，频数为 1；落在第二组 64.5～69.5 内的瓶数有 66，67，68，68 共四个，频数为 4。第四列为频率，即相对频数。例如，第一组的频数为 1，总频数为 40，其频率为 1/40 = 0.025；第二组频数为 4，故相对频数为 4/40 = 0.100。

表1-2 某商店日销售啤酒瓶数频率分布

组距(上下限)	组中心	频数	频率(相对频数)
59.5～64.5	62	1	0.025
64.5～69.5	67	4	0.100
69.5～74.5	72	8	0.200
74.5～79.5	77	11	0.275
79.5～84.5	82	6	0.150
84.5～89.5	87	7	0.175
89.5～94.5	92	3	0.075

二、直方图与频数多边形

表 1-2 的数据可以表示成直方图，如图 1-2 所示。方法是：把组距作为横坐标，频数作为纵坐标，在每一组上竖起一个长方形。这样的图形就称为直方图。把每个长方形顶边的中点用折线连结起来，所围成的多边形，就称为频数多边形。

三、累积频率分布

在许多情况下，我们需要知道小于或等于某一数量的频数。例如，我们要求知道啤酒日销

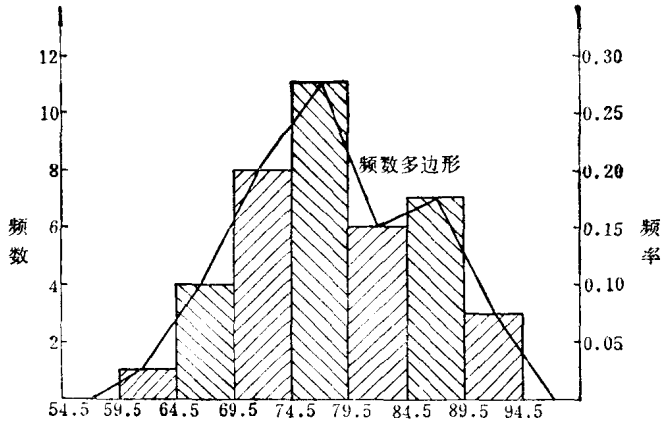

图1-2　直方图与频数多边形

售瓶数小于 79.5 的瓶数。根据表 1-2，我们只需把第一列中小于 79.5 的四个组的对应频数即第三列中的 1，4，8，11 相加，便可算出小于 79.5 的频数为 24，这个频数就叫做累积频数。同理，把这四个组所对应的频率相加，所得结果就叫做累积频率。表 1-3 便是啤酒日销售量的累积频数和累积频率表。

表1-3　啤酒日销售量累积频数和累积频率

组　　距	组　中　心	频　　数	累　积　频　数	累　积　频　率
59.5～64.5	62	1	1	0.025
64.5～69.5	67	4	5	0.125
69.5～74.5	72	8	13	0.325
74.5～79.5	77	11	24	0.600
79.5～84.5	82	6	30	0.750
84.5～89.5	87	7	37	0.925
89.5～94.5	92	3	40	1.00

以组距为横坐标，累积频率为纵坐标，在每个组距上竖起一个长方形，其底长等于组距，高等于累积频率，我们就可得到表 1-3 所列数据的图形，叫累积频率分布图(见图1-3)。在图 1-3 中，用折线连结每个长方形顶边的中点，就得到了累积频率多边形图。

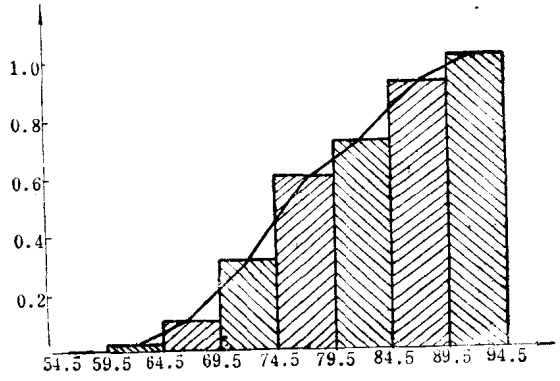

图1-3　累积频率分布图

§1-4 中心位置特征值：平均数、中位数和众数

我们在上一节中把数据组织成表格和图形以后，可以看到，这些图表已经显示了研究对象的某种规律性的面貌；但是如果我们想要借助这些图表记住或比较分析这些数据，那还是一件比较烦琐的事情。因此，我们还需要一些概括性的量度——分布的特征值，它们能够以数量的形式表示这些数据的特征。概括性的特征值主要有两类：一类表示数据的中心位置；另一类表示数据的变异程度，或离散程度。两者互为补充，概括反映数据的全貌。

一、平均数

表示数据中心位置的一个主要特征值就是数据的算术平均值，简称平均值或均值。例如，有一组数据共有 n 个，为：

$$x_1,\ x_2,\ x_3,\ \cdots,\ x_n$$

其平均数 $\bar{x}$ 就是：

$$\bar{x}=\frac{1}{n}\sum_{i=1}^{n}x_i \tag{1-1}$$

表1-4 平均数、中位数、众数

	甲组数据	乙组数据	丙组数据	丁组数据
数据	2、2、3、4、8、10、13	5、8、8、9、10、12	2、3、4、4、4、7	2、3、4、4、4、19
平均数	6	7	4	6
中位数	4	8.5	4	4
众数	2	8	4	4

表 1-4 列出了四组数据，根据公式 (1-1)，我们可求得甲组数据的平均数：

$$\bar{x}=\frac{2+2+3+4+8+10+13}{7}=\frac{42}{7}=6$$

二、中位数

表示数据中心位置的第二个特征值叫做中位数。确定中位数，首先要把数据按大小次序排列起来，然后取居于中间的一个，即为中位数，记为 $\tilde{x}$。什么叫做中间的一个呢？即在它前面的数据的个数等于在它后面的数据的个数。如果一组数据共有 N 个，且 N 为奇数，则中位数便是第 $\frac{N+1}{2}$ 个。例如，在表 1-4 中，甲组数据共有 7 个，$N=7$，是奇数，故其中位数是第 $\frac{7+1}{2}=4$，即第四个，正好等于 4。若 N 为偶数，则中间便有两个数据，其中位数定义为此两个中间数据之和的一半。例如，在表 1-4 中，乙组共有 6 个数据，$N=6$，为偶数，其中间的两个数据为 8 和 9，因此，它的中位数等于 $\frac{8+9}{2}=8.5$。

三、众数

表示数据中心位置的另一个特征值是众数。它就是一组数据中频数最多的那个数据。例如在表 1-4 中，甲组数据的众数为 2，因为 2 的频数在这组数据中是最多的。同理，不难看

出乙、丙、丁三组数据的众数分别为8、4、4，因为它们都是本组数据中频数最多的一个。

在以上三个表示数据中心位置的特征值中，平均数是广泛采用的一个位置中心量度。与中位数和众数相比，它容易受到数据极端值的影响。例如在表1-4中，丙、丁两组数据仅有一个数不同，丁组最后一个数据为19，丙组最后一个数据为7。两组数据的中位数和众数未受到此两数不同的影响，都等于4，但两组数据的平均值却不同，丙组为4，丁组为6，受到了极端值的影响。这是平均数的缺点。在要求数据只作大小排列而以数据顺序为主要研究特征时，中位数是理想的中心位置特征值。在只需研究发生频次最多的情况时，众值是最佳的中心位置特征值。当既需要了解数据的大小顺序、发生频数以外，还要求注意数据值的大小时，平均数便是一个比较全面的特征值，因为它充分利用了这三个方面的信息。例如在表1-4中，甲组数据的众值为2，正好是最小的一个数，作为甲组数据的中心位置显然是不恰当的。这正是众数的缺点，因为它只照顾了数据发生的频数。由此可见，三个数据中心位置特征值各有不同的特点，运用时必须根据具体情况选择确定。

四、几何平均数与调和平均数

除了用算术平均数、中位数和众数三个特征值作为数据中心位置的量度以外，有时所得到的数据是一种比率，所要求的平均数是平均比率。这时，就需要采用几何平均方法。

$$G = \sqrt[n]{x_1 \cdot x_2 \cdot x_3 \cdots x_n} \tag{1-2}$$

式中的G为几何平均数，x_1，x_2，$x_3 \cdots x_n$为数据。

例如，某商品在四年期间的年初价格分别为：

30.00元；33.00元；33.66元；41.74元

每年初与上一年初价格的比率为：

$$1.10\left(=\frac{33.00}{30.00}\right);\ 1.02\left(=\frac{33.66}{33.00}\right);\ 1.24\left(=\frac{41.74}{33.66}\right)$$

则其每年价格与上年价格的平均比率应为：

$$G = \sqrt[3]{1.10 \times 1.02 \times 1.24} = 1.116$$

这样，按第一年初的价格30.00元为基数，第四年初的价格便为

$$30.00 \times 1.116 \times 1.116 \times 1.116 \approx 41.74 \quad 元$$

有时，所得到的数据是一组变化率。例如，一辆汽车在公路上行驶，在第一段10km路上，以50km/h速度行驶，在第二段10km路上，因路面狭窄，降速行驶，速度为30km/h。如果欲求这辆汽车的行驶平均速度，就需要使用调和平均数。第一段10km路程花了$10 \times \frac{1}{50} = 0.2$h，第二段10km路程花了$10 \times \frac{1}{30} = 0.333$ h，合计20km路程共花了0.533 h，所以这辆汽车的平均速度应是20/0.533 = 37.5km/h。这个数字就是一个调和平均数。

$$H = \frac{n}{\frac{1}{x_1} + \frac{1}{x_2} + \frac{1}{x_3} + \cdots + \frac{1}{x_n}} \tag{1-3}$$

式中的H为调和平均数，x_1，x_2，…，x_n为n个表示变化率的数据。

就上例来讲，按(1-3)式，有平均速度：

$$H=\frac{2}{\frac{1}{50}+\frac{1}{30}}=\frac{2}{0.02+0.033}$$

$$=\frac{2}{0.05333}=37.5\quad(\mathrm{km/h})$$

§1-5 变 异 程 度

数据的变异是统计的一个特征。数据变异程度的大小反映数据的特性。例如测量人的血压，就需要有上、下限两个数据：80～120mmHg⊖是正常状态，90～180mmHg是高血压的状态。在这里，上下两个数据的平均数就无多大意义了。

在§1-4中我们阐述过反映数据中心位置的特征值。由于数据存在变异性，可以设想，具有同一个平均数的两组数据可以完全不相同。例如，6、7、8三个数的平均数等于7；另一组数据0，7，14的平均数亦是7。但是，在这两组数据中，除了7以外，其它四个数完全不相同，而且后一组数据的变化范围显然要比前一组数据大得多。所以，仅仅用中心位置特征值来描述数据的分布是不够全面的，还需要有一个描述数据变异程度或离散程度的概括性量度，来补充位置特征值的不足。反过来说，仅用变异程度这个特征值来描述数据也是不够全面的，还必须和中心位置特征值结合起来，才能较好地显示数据的全貌。

一、极差

数据的最大值和最小值的差，叫做极差。它是描述数据变异程度的简便量度。如以x_{max}和x_{min}分别表示数据的最大值和最小值，R表示极差，则：

$$R=x_{max}-x_{min}$$

例如，表1-1中啤酒日销售数的极差是：

$$R=94-63=31$$

二、中间位差

极差在一定程度上固然描述了数据的变异程度，但它忽视了中间所有的数据，并要受极端值的过大影响，所以它不是对变异程度理想的量度。为了避免极端值的过大影响，可以排除两端的一些数据，只取中间一部分数据的极差。这部分数据的极差就叫做中间位差。中间位差既无极差的缺点，又能表示数据的变异程度。中间位差以其所包含数据的百分率命名。例如，80%中间位差就是指除去前面及后面各10%的数据后，所余中间数据的极差。以表1-1中所列的数据为例，其

$$80\%\text{中间位差}=84-73=11$$

它就是前后除去各10个数后所得的极差。

三、平均离差

设有n个数据

$$x_1,\ x_2,\ \cdots,\ x_n$$

其平均数为$\bar{x}$。现将每个数$x_i(i=1,2,\cdots,n)$与平均数$\bar{x}$相减，求出其绝对差值，即：

$$|x_i-\bar{x}|$$

⊖ 1mmHg=133.332Pa，下同。

再将 n 个绝对差值进行平均

$$\frac{\sum_{i=1}^{n}|x_i-\bar{x}|}{n} \tag{1-4}$$

所得结果就叫做这组数据的平均离差。平均离差表示各个数据离开它们的中心位置的平均距离，故亦为测定数据离散程度的指标之一。

四、方差与标准差

比较理想的描述数据变异程度的两个特征值是方差和标准差。设一个样本的 n 个数据为

$$x_1, x_2, \cdots, x_n$$

求出它们和平均数 $\bar{x}$ 的离差平方和，再除以 $n-1$，就得到

$$s^2=\frac{\sum_{i=1}^{n}(x_i-\bar{x})^2}{n-1} \tag{1-5}$$

叫做方差（样本方差）。对方差开方取正值，可得标准差⊖

$$s=\sqrt{\frac{\sum_{i=1}^{n}(x_i-\bar{x})^2}{n-1}} \tag{1-6}$$

标准差作为描述数据离散程度的一个特征值，其优点首先在于它与数据的中心位置无关（即不受指标平移的影响）；其次，由于对数据与平均数的离差取平方，可以排除正负离差相互抵销的问题；其三，它充分利用了每个数据所提供的信息；最后，它在数学上还容易处理。

计算方差 s^2 的一个等价公式是：

$$s^2=\frac{\sum_{i=1}^{n}x_i^2-\left(\sum_{i=1}^{n}x_i\right)^2\Big/n}{n-1} \tag{1-7}$$

这是因为：

$$\sum_{i=1}^{n}(x_i-\bar{x})^2=\sum_{i=1}^{n}(x_i^2-2\bar{x}x_i+\bar{x}^2)$$

⊖ 有的教材用 n 作分母，方差和标准差分别是：

$$s^2=\frac{\sum_{i=1}^{n}(x_i-\bar{x})^2}{n}$$

$$s=\sqrt{\frac{\sum_{i=1}^{n}(x_i-\bar{x})^2}{n}}$$

这里我们采用的分母为 $n-1$，计算出来的 s^2 为 σ^2 的无偏估计量。对此，将在第三章中加以解释。

$$= \sum_{i=1}^{n} x_i^2 - 2\bar{x}\sum_{i=1}^{n} x_i + n\bar{x}^2$$

$$= \sum_{i=1}^{n} x_i^2 - \left(\sum_{i=1}^{n} x_i\right)^2 \Big/ n$$

将此结果代入(1-5)式，便得到了(1－7)式。同理，可得到计算 s 的一个等价公式：

$$s = \sqrt{\frac{\sum_{i=1}^{n} x_i^2 - \left(\sum_{i=1}^{n} x_i\right)^2 \Big/ n}{n-1}} \tag{1-8}$$

可以看出，标准差 s 是 s^2 的算术根。两者在本质上反映的是同一现象，其差别是 s 的单位与变量是一致的，而 s^2 的单位则是变量单位的二次方。

五、总体平均数与总体标准差（μ 与 σ）

我们在§1-1中曾谈到所研究的一组数据有总体和样本两种不同的情况。以上我们讨论的一些特征值，都假定数据是一个样本。这些样本特征值都叫做统计量。特别是平均数 $\bar{x}$ 与标准差 s，将是本书以后将要经常讲到的两个重要的统计量。现在我们应该指出，如果所有数据表示的是一个总体，那末与 $\bar{x}$ 和 s 相对应的特征值就叫做总体的参数。它们分别是：

$$\mu = \frac{\sum_{i=1}^{N} x_i}{N} \tag{1-9}$$

与

$$\sigma = \sqrt{\frac{\sum_{i=1}^{N} (x_i - \mu)^2}{N}} \tag{1-10}$$

其意义也是平均数与标准差。由于它们来自总体，故其符号不同，计算公式也略有不同。式中的 N 代表总体数据的个数；式中的差值是数据与总体平均数 μ 相减计算出来的；分母为 N，而不是 $N-1$，这一点应该特别加以注意。

§1-6 分组数据特征值的计算

上述求平均数与标准差的公式〔公式(1-1)，(1-8)，(1-9)，(1-10)〕，都是按数据分别给出的数值进行计算的。在已将数据整理组成频数表时，便可按组分别取组中值并乘上各该组的频数，求数据的平均数，或用类似的方法求数据的标准差。因此，对应于公式(1-1)，(1-8)，(1-9)，(1-10)，有：

$$\bar{x} = \frac{\sum_{i=1}^{c} f_i x_i}{\sum_{i=1}^{c} f_i} \tag{1-11}$$