药学应用概率统计

Yaoxue Yingyong Gailu Tongji

韩可勤 杨静化 刘晓东 编

东南大学出版社

薬学応用数理統計

药学应用概率统计

韩可勤　杨静化　刘晓东　编

东南大学出版社

·南　京·

内容提要

本书介绍了概率统计的基本内容、生物统计方法和SAS统计分析。鉴于医药院校学生的专业特点，本书力求叙述清楚，简明易懂，应用性强。各章配有大量的例题和习题，便于教学和自学。

本书可用作医药院校各专业学生教材，也可供医药卫生工作者自学或参考。

图书在版编目(CIP)数据

药学应用概率统计/韩可勤，杨静化，刘晓东编著．
南京：东南大学出版社，2000.9
ISBN 7-81050-658-1

Ⅰ.药… Ⅱ.①韩…②杨…③刘… Ⅲ.①概率论-应用-药物学-研究②数理统计：生物统计-应用-药物学-研究

中国版本图书馆 CIP 数据核字(2000)第 36243 号

东南大学出版社出版发行
（南京四牌楼 2 号 邮编 210096）
出版人：宋增民
江苏省新华书店经销 中国药科大学印刷厂印刷
开本：787mm×1092mm 1/16 印张：17.5 字数 437 千字
2000 年 9 月第 1 版 2000 年 9 月第 1 次印刷
印数：1-5040 册 定价：22.00 元

前　言

概率统计是研究随机现象的数量规律的一门科学，在医药学领域中有着非常广泛的应用，是医药工作者必备的基础知识。针对医药领域各专业的特点和需要，结合医药应用的实际背景，本书编者力图把概率统计的基本概念、基本原理讲解清晰，循序渐进。为帮助读者易于理解，尽量多作直观解释，辅以医药应用实例进行说明，避免较长的数学证明。在充分注意系统性和可接受性的同时，本书突出其应用性。为有利于读者运用统计方法解决现代药物研究和生产的实际问题，编者专门写了生物统计方法一章，以介绍统计方法在药学上的一些应用，并尝试把目前国际上医药领域中广为使用的SAS统计分析软件融入教材之中，对常用的统计方法都配有SAS应用实例。此外，书中还配有大量带有应用性的例题与习题。

本书含基础概率和常用统计方法两部分内容。具体为：统计资料的描述、概率及其计算、随机变量的分布和数字特征、抽样分布和参数估计、假设检验和方差分析、正交设计与均匀设计、相关分析与回归分析、生物统计方法等。其中以介绍统计方法的应用为主，注意讲明各种方法的背景、应用条件及实际意义。讲授全书内容约需72学时。

本书前八章由韩可勤编写，第九章由刘晓东编写，第十章由杨静化编写。另外，高祖新老师和丁峻老师参加了本书的编写讨论会，倪永兴老师审阅了部分书稿，他们对本书内容的取舍和体系结构都提出了许多有益的建议，为本书增色不少，特此表示感谢。

本书除可供医药院校各专业作教材使用外，还可供医药卫生工作者自学或参考。

限于水平，书中不妥之处在所难免，敬请读者批评指正。

编　者

于中国药科大学

1999年12月

目 录

第一章　统计资料的描述

统计(Statistics)是一门关于数量资料的搜集、整理、分析和解释的科学，在药学领域中有着广泛的应用。药学工作者和药政管理人员在新药研制、药物鉴定、试验设计、药政管理、处方优选等许多方面，面临着大量的数量资料(Data)有待整理，分析。因此，学习有关的统计知识和接受必要的统计训练，对指导如何有效利用数据资料进行统计实践十分有益。

§1　统计资料的整理

一、统计资料的类型

药学统计数据一般分为计量资料和计数资料，介于其中的还有等级资料，不同类型的资料应采用不同的分析方法。

1. **计量资料**(Measurement data)　对每个观察单位用定量方法测定某项指标量的大小，所得资料称为计量资料。例如调查某地12岁男童的身体发育状况，以人为观察单位，每个人的身高(cm)、体重(kg)和血压(mmHg)* 等；又如以每个采样点为观察单位，测得不同采样点的二氧化碳浓度(mg/L)等。这类资料一般具有计量单位、各观察单位的测量值常有量的差异。分析计量资料常用假设检验、方差分析等。

2. **计数资料**(Enumeration data)　将观察单位按某种属性或类别分组，所得各组的观察单位数，称为计数资料。例如测试某班学生仰卧起坐次数；用某药治疗若干流感病人后的治愈人数和未愈人数；某人群中O、A、B、AB各种血型的人数。分属于各组的观察单位间有质的差别，不同质的观察单位不能归在同一个组内。分析计数资料常用 χ^2 检验等。

3. **等级资料**(Ranked data)　将观察单位按某种属性的不同程度分组，所得各组的观察单位数，称为等级资料。例如用某药治疗若干痢疾病人，其中治愈、好转、无效人数；评定新药研制水平的高低分为一类、二类、三类、四类等。这类资料与计数资料不同的是：属性的分组有程度的差别，各组按大小顺序排列；与计量资料不同的是：每个观察单位未确切定量，因而称为半计量资料。分析等级资料常用秩和检验等。

根据分析的需要，计量资料、计数资料和等级资料可以互相转化。例如每个人的血红蛋白，原属计量资料，若按血红蛋白正常与异常分为两组，得各组人数，是计数资料。若将血红蛋白按量(g/L)的多少分为五个等级：小于60(重度贫血)、60～(中度贫血)、90～(轻度贫血)、120～160(血红蛋白正常)、大于160(血红蛋白增高)，得各等级人数，就是等级资料。

二、统计资料的整理

信息的搜集往往得出大量的统计资料，要使这些资料能被人们看懂，从中提取出有用的

---

* 1mmHg＝133.322Pa

信息，进而探索其中的数量规律性。因此，统计资料必须进行整理。**统计整理** 是以适当形式展示已收集的资料，以便得出合乎逻辑的结论。将统计资料分组是数据整理的基本方法。统计分组是按不同的标志特征把数据划分为性质不同的若干类别或几个部分。例如，学生按性别、地区等分类称为按**品质标志分组** ，按年龄、成绩等分类称为按**数量标志分组** 。数据观察值在各组中的个数称为**次数**（也称**频数** ），各组间的次数就是频数分布。展示数量资料的常用方式是统计表和统计图。把记录各数据观察值出现的次数作成表格形式，就是频率分布表。为直观、清晰地展示频数分布，常用直方图表示。下面将通过一个具体实例介绍数据整理的过程。

例 1　为研究某地区 12 岁男孩身高的分布情况，随机地抽取 120 名男孩，测得身高数据如下（单位：cm）：

128.1	144.4	150.3	146.2	140.6	139.7
134.1	124.3	147.9	143.0	143.1	142.7
126.0	125.6	127.7	154.4	142.7	141.2
133.4	131.0	126.4	130.3	146.3	146.8
142.7	137.6	136.9	122.7	131.8	147.7
135.8	134.8	139.1	139.0	132.3	134.7
138.4	136.6	136.2	141.6	141.0	138.4
145.1	141.4	139.9	140.6	140.2	131.0
150.4	142.7	144.3	136.4	134.5	132.3
152.7	148.1	139.6	138.9	136.1	135.9
140.3	137.3	134.6	145.2	128.2	135.9
140.2	136.6	139.6	135.7	139.8	129.1
141.4	139.7	136.2	138.4	138.1	132.9
142.9	144.7	138.8	138.3	135.3	140.6
142.6	152.1	142.4	142.7	136.2	135.0
154.3	147.9	141.3	143.8	138.1	139.7
127.4	146.0	155.8	141.2	146.4	139.4
140.8	127.7	150.7	160.3	148.5	147.5
138.9	123.1	126.0	150.0	143.7	156.9
133.1	142.8	136.8	133.1	144.5	142.4

试编制频数分布表。

解：上表为 120 名男孩身高的原始数据，如不加以整理很难发现它的规律性。现用此例来说明数据整理的一般方法。

（1）找出数据中的最大值 X_{max} 和最小值 X_{min}，并确定极差（Range）。最大值与最小值之差称为**极差** 或**全距** ，常用 R 表示。本例 $X_{max}=160.3$，　$X_{min}=122.7$，　$R=160.3-122.7=37.6$。

（2）确定组距 d 和组数 k。大多按等距分组，有时也按不等距分组。分组时，组数要适当，因为组数太多，有时组内可能没有样本数据；组数太少，不易看出分布的特征。

一般地说，当数据个数小于50时，可分为5～6组；当数据个数为100左右时，可分为7～10 组；当数据个数更大时，可分为 10～15 组，或根据下列组数估计公式来确定组数：

$$k=1.87(N-1)^{2/5}$$

比如，当 $N=200$ 时，$k=16$；$N=500$ 时，$k=22$。

组距取决于组数和极差：

$$d=\frac{R}{k}$$

本例 $N=120$，可分为 10 组，极差 $R=37.6$，故组距 d 宜取 4cm。这样，用分点

$$a=x_0<x_1<x_2<\cdots<x_{10}=b$$

将区间(a,b)分成 10 个小区间，始点 a 为 122.0cm，终点 b 为 162.0cm，每一小区间均为半开半闭区间，以免重叠(表 1.1)。

(3)统计频数，算出频率和频率密度。用唱票方法统计出落在各个小区间内的数据个数，即频数 n_i，各组的频率为 $f_i=\frac{n_i}{N}$。最后，将各组的频率 f_i 除以相应的组距 d，便得到频率密度 $\frac{f_i}{d}$(表 1.1)。

表 1.1　频率分布表

组　号	区　　间	频　数　划　记	频　数	频　率	频率密度
1	[122.0,126.0)	丅	4	0.033	0.00825
2	[126.0,130.0)	正丅	9	0.075	0.01875
3	[130.0,134.0)	正正	10	0.083	0.02075
4	[134.0,138.0)	正正正正丅	22	0.183	0.04575
5	[138.0,142.0)	正正正正正正丅	33	0.275	0.06875
6	[142.0,146.0)	正正正正	20	0.167	0.04175
7	[146.0,150.0)	正正一	11	0.092	0.02300
8	[150.0,154.0)	正一	6	0.050	0.01250
9	[154.0,158.0)	丅	4	0.033	0.00825
10	[158.0,162.0)	一	1	0.008	0.00200
合　计			120	0.999	

频数分布的用途如下：

(1)揭示频数分布特征。从表 1.1 中可看出频数分布有两个重要特征：一是 12 岁男童身高虽有高有矮，但过高或过矮的人是少数，而中等身材的居多，称为集中趋势；另一是身高参差不齐，称为离散趋势。本章后面将进一步讨论测定集中趋势和离散趋势的方法，以便全面地认识和分析被研究的事物。

(2)便于发现某些特大或特小的异常值。对异常值应进一步检验，是否有测定上的差错，核对原始数据，进而经过统计判断，决定取舍。

(3)提供分组数据，以便进一步计算和分析。

三、直方图和累积频率函数图

频数分布也可以用各种图形表示，使得频数分布的数量规律性更直观、更形象。常用的图形有直方图、频数曲线图和累积频率函数图等。

1. 直方图 (Histogram)

直方图是由若干矩形所组成的，每个矩形的宽等于组距，而高是各组的频数或频率。用表 1.1 的资料绘制的直方图见图 1.1。

以变量取值为横坐标，频率密度为纵坐标，以组距 d 为底，频率密度 $\frac{f_i}{d}$ 为高，在每个小区

间上作出小矩形(矩形的面积为频率),便得直方图。最后,按直方图的分布特点适当画出一条光滑曲线(图 1.1)。由图 1.1 可见,曲线呈中间高两头低,左右近似对称,与正态曲线相像。

2. 累积频率图

由第 1 组起到第 i 组止各频率之和称之为累积频率。频率分布和累积频率分布由于不受总数 N 的影响,便于不同资料的比较。

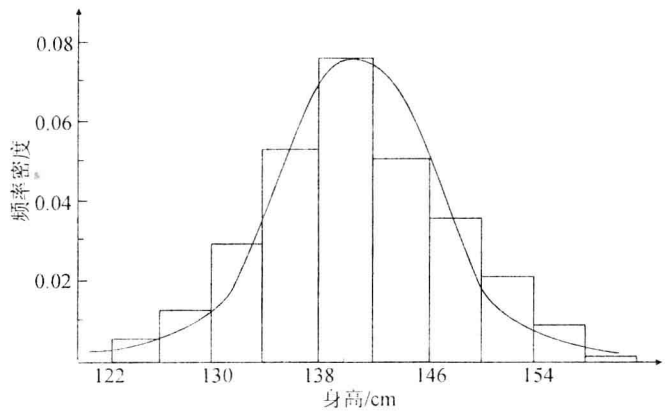

图 1.1 频率直方图

下面利用表 1.1 的数据来介绍累积频率图的作法。首先,由表 1.1 列出表 1.2,然后用一条光滑的曲线连接,即得累积频率函数图(图 1.2)。

表 1.2 累积频率表

组号	组中值	频率	累积频率
1	124.0	0.033	0.033
2	128.0	0.075	0.108
3	132.0	0.083	0.191
4	136.0	0.183	0.374
5	140.0	0.275	0.649
6	144.0	0.167	0.816
7	148.0	0.092	0.908
8	152.0	0.050	0.958
9	156.0	0.033	0.991
10	160.0	0.008	0.999

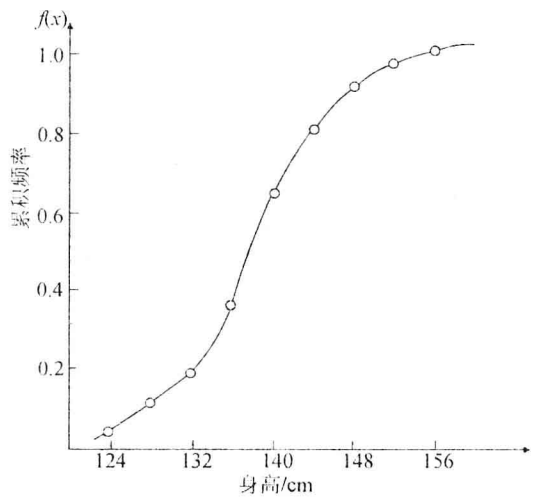

图 1.2 累积频率函数图

对例 1 中的数据资料可借助 SAS 统计分析软件进一步分析,如用茎叶法等。对统计数据的整理,前面所介绍的是先将数据分组,然后形成次数分配的传统方法。茎叶法可将统计分组和次数分配两项工作一次完成,图形直观且保留原始信息。在计算机日益普及的今天,此方法更显示出其优越性。例 1 的茎叶法等分析结果见第十章 § 4。

四、总体与样本

总体 (Population)是根据研究目的确定的同质的探讨对象的全体,组成总体的每一个单元称为**个体** (Individual)。例如,要研究某一个工艺条件下生产的一批针剂,其总体是这批针剂,每一支针剂都是这个总体的一个个体。又如研究某地 12 岁儿童的健康状况,总体就是全体 12 岁儿童,每一个 12 岁儿童为个体。不过我们通常仅仅关心研究对象的某种指标的情况。如对针剂,我们希望知道它们的有效期、药物含量等;对 12 岁儿童的发育情况,我们希望

了解他们的身长、体重等等。这些不同数值，反映现象特征的指标，也称为**变量**。统计中，总体也是指某变量 X 取值的集合，或简单地说成总体 X。

总体视其所含个体的个数是有限的还是无限的，又分为有限总体和无限总体。即使对有限总体，若个体数过多，直接研究总体仍费时费力，有时是不可能的和不必要的。所以在实际工作中，常是在总体中随机抽取部分个体构成**样本** (Sample)，用样本信息来推断总体特征。如例 1 中，研究某地 12 岁男孩身高的分布情况，其研究对象是该地 12 岁所有男孩，即总体。随机抽取 120 个男孩为样本。在例 1 中，通过对 120 名男孩身高(样本资料)的研究来推断该地 12 岁男孩身高的总体情况。样本中所含个体的个数称为**样本容量**。例 1 的样本容量为 120。根据样本资料作出的直方图、累积频率函数图亦称为样本直方图、样本累积频率函数图。

§2 集中趋势的测度

原始数据经过分组整理后，形成了频数分布。将频数分布用直方图画出来，我们对该组数据的变化规律就有了直观的了解。然而，进一步的推断、决策等研究不仅要求我们对其分布的规律有直观的了解，而且要求我们用几个最简洁又最能充分描述其分布数量特征的统计量将其分布变化的规律性表示出来。表示集中趋势的数字特征有平均数、中位数、众数等。

一、算术平均数(Arithmetic average)

算术平均数又称为均值，用 $\bar{x}$ 表示。

例 1 两药厂生产同一种产品、每天分别各抽取 100 件进行检查，连续抽查 10 天，其次品数如表 1.3 所示。

表 1.3 甲、乙两厂次品数统计

甲厂(x_1)	9	5	1	6	5	4	5	5	4	6
乙厂(x_2)	0	10	2	9	7	9	0	10	10	2

试比较两厂产品的优劣。

解：很显然，每天的次品数多，质量就差，但两厂每天抽检到的次品数有高有低，所以应该比较平均每天抽检到的次品数。即用算术平均数可以表示甲、乙两厂次品数的集中趋势。

以 $\bar{x}_1$，$\bar{x}_2$ 分别表示两厂平均每天抽检到的次品数，则

$$\bar{x}_1=\frac{1}{10}(9+5+1+6+5+4+5+5+4+6)=5$$

$$\bar{x}_2=\frac{1}{10}(0+10+2+9+7+9+0+10+10+2)=5.9$$

可见，甲厂平均每天抽检到次品数小于乙厂。平均数越大，代表整个厂次品数越多。因此，平均数是各厂日次品数的集中性表征。一般平均数的定义如下：

1. 定义：一组 n 个观察值 $x_1,x_2,\cdots,x_n$ 的算术平均数 $\bar{x}$ 为

$$\bar{x}=\sum_{i=1}^{n}x_i/n \tag{1.1}$$

如果资料已经分组，组数为 k，用 $x_1,x_2,\cdots,x_k$ 表示各组中点，$f_1,f_2,\cdots,f_k$ 表示相应的频

数，那么

$$\bar{x}=\sum_{i=1}^{k}f_i x_i/\sum_{i=1}^{k}f_i \tag{1.2}$$

或

$$\bar{x}=\sum_{i=1}^{k}\left(x_i\cdot\frac{f_i}{\sum_{i=1}^{k}f_i}\right) \tag{1.3}$$

公式(1.1)称为平均数的简单式，式(1.2)和式(1.3)称为加权式，公式(1.3)表明均值不仅受到变量值 x_i 大小的影响，而且还受到权重系数 $\frac{f_i}{\sum_{i=1}^{k}f_i}$ 大小的影响。

利用公式(1.3)可由表 1.1 计算出 120 名男孩的平均身高为 139.73cm，此结果只是个近似的值。在求平均数时应尽可能用分组前的原始数据，如利用公式(1.1)求出上节例 1 中的平均值才是精确值。

2. 均值的数学性质

(1) $\sum_{i=1}^{k}(x_i-\bar{x})=0$ 或 $\sum_{i=1}^{k}(x_i-\bar{x})f_i=0$

即数据观察值与均值的离差之和为零。

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

A 为任一不等于 $\bar{x}$ 的数值，即均值的离差平方和最小。该性质是以后要讲到的最小二乘法的基础。

二、中位数和众数

算术平均数表示了集中趋势的特征，它照顾到每一个值，其缺点就是受观测值中极端情况的影响很大。例如，某班 5 名学生的成绩为 93,90,85,82,0，其平均成绩 $\bar{x}=70$ 就没有什么代表性，因此有必要研究表示集中趋势的其他数字特征。

1. 中位数 (Median)

定义：一组 n 个观察值按数值由小到大顺序排列后为 $x_1,x_2,\cdots,x_n$。

处于中间位置的值称为中位数，用 Me 表示，即

$$Me=\begin{cases}x_{\frac{n+1}{2}} & \text{当 } n \text{ 为奇数}\\ \frac{1}{2}\left(x_{\frac{n}{2}}+x_{\frac{n}{2}+1}\right) & \text{当 } n \text{ 为偶数}\end{cases} \tag{1.4}$$

对于已分组的频数分布，只能求中位数所在的组，即累积频数含 $n/2$(或累积频率含 0.50)的组。例如，在表 1.2 累积频率 0.5 的组为第 5 组，即 138～142，为中位数所在组。

2. 众数 (Mode)

众数 是一组数据中出现最多的那个观察值，用 Mo 表示。

对分组且等距的频数分布，根据最大频数，只能求得众数所在组，不能求得众数的确切值。由表 1.1 可知上节例 1 中最大频数为 33，于是知众数所在组为第 5 组，即区间为 [138.0,142.0)。

三、$\bar{x}$, Me, Mo 三者位置关系

比较众数 Mo，中位数 Me，算术平均数 $\bar{x}$ 的相对位置关系，可以研究频数分布的偏倚性。所谓**偏倚性** 是表示各观察值分布的不对称程度的指标。图 1.3 表示了对称、**左偏**（负偏）和**右偏**（正偏）的频数分布例子。注意到它们的特点是：(1)对称分布的众数、中位数和算术平均数相同；(2)具有偏倚性的分布，算术平均数突出在外，偏向分布的尾端，而中位数则介于众数与算术平均数之间。

即对于单峰分布：

对称：$\bar{x}=Me=Mo$

左偏：$\bar{x}<Me<Mo$

右偏：$\bar{x}>Me>Mo$

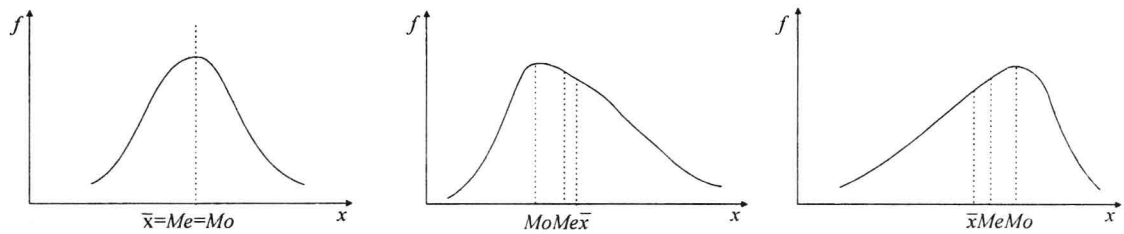

图 1.3　不同分布下 $\bar{x}$、Me、Mo 关系图

§3　离散程度的测度

前一节讨论了对一组数据集中趋势测度的方法。要把握一组数据的数量变化规律仅仅有集中趋势的测度是不够的，还要了解数据的离散程度。如果这组数据是测量的结果，那么离散程度说明测量方法或仪器是精密还是粗糙；如果数据是产品质量检验结果，那么数据的离散情况说明生产是否稳定。

例 1　甲、乙两名化验员分析同一样品各 5 次，所得结果为

甲：	5.2	5.1	5.0	4.9	4.8
乙：	6.0	5.5	5.0	4.5	4.0

两者的平均数都是 5.0，能否认为两人的技术水平相同？

只要稍为留意一下就会发现，甲的各次结果比较接近平均数，而乙的各次结果比较离散。

统计学中测度离散程度应用最多的指标是方差和标准差。

一、方差(Variance)和标准差(Standard deviation)

方差 是观察值与其均值离差的平方和的均值，它又有总体方差(Population variance)和样本方差(Sample variance)之分。

如果观察数据是总体数据，则总体方差的计算公式为

$$\sigma^2 = \sum_{i=1}^{N}(x_i - \bar{x})^2/N \tag{1.5}$$

标准差是方差的正平方根，即

$$\sigma = \sqrt{\frac{\sum_{i=1}^{N}(x_i - \bar{x})^2}{N}} \tag{1.6}$$

如果观察值是一组样本数据，则样本方差的计算公式为

$$S^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2 \tag{1.7}$$

式中，$n-1$ 为**自由度**。自由度是统计中的常用术语，其意义以后讨论。

样本标准差是样本方差的正平方根，即

$$S = \sqrt{\frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n-1}} \tag{1.8}$$

对于已分组的频数分布(组数为 k)

$$S^2 = \frac{\sum_{i=1}^{k}(x_i - \bar{x})^2 \cdot f_i}{\sum_{i=1}^{k} f_i - 1} \tag{1.9}$$

样本标准误 $S_{\bar{x}}$为

$$S_{\bar{x}} = \sqrt{\frac{1}{n(n-1)}\sum_{i=1}^{n}(x_i - \bar{x})^2} \tag{1.10}$$

实际计算中，有时用样本方差的另一种表达式：

$$S^2 = \frac{1}{n-1}\left[\sum_{i=1}^{n}x_i^2 - (\sum_{i=1}^{n}x_i)^2/n\right] \tag{1.11}$$

我们把样本均值 $\bar{x}$与样本方差 S^2 统称为样本的数字特征。此在后面的统计推断中有重要的应用。

例 1 中的数据假定为样本数据，可分别计算出：$S_{甲}^2 = 0.025$，$S_{乙}^2 = 0.625$，由于甲、乙两组数据的平均数相同，$\bar{x}_{甲} = \bar{x}_{乙} = 5$，而 $S_{甲}^2 < S_{乙}^2$。因此，化验员甲的技术水平较为稳定。

例 2　某医院测得矽肺病人治疗前血中粘蛋白含量(mg%)为 6.5，7.3，3.0，7.3，5.6，6.2，7.3，求样本的均数、方差、标准差、标准误、中位数、众数和极差。

解：易得

$$\sum x_i = 43.2 \qquad \sum x_i^2 = 280.92$$

$$\bar{x} = \frac{1}{7}\sum x_i = \frac{1}{7} \times 43.2 \approx 6.17$$

于是

$$S^2 = \frac{1}{6}(\sum x_i^2 - \frac{1}{7}(\sum x_i)^2) \approx 2.3857$$

$$S = \sqrt{2.3857} \approx 1.54$$

$$S_{\bar{x}}=\frac{S}{\sqrt{7}}=\frac{1.54}{\sqrt{7}}=0.58$$

将样本值按大小顺序排列，有 3.0,5.6,6.2,6.5,7.3,7.3,7.3

则得

$$Me=6.5 \qquad Mo=7.3$$

$$R=x_{max}-x_{min}=7.3-3.0=4.3$$

二、变异系数(Coefficient of variation)

$$CV=\frac{S}{\bar{x}}\times 100\% \tag{1.12}$$

CV 是一个无量纲的量，它适用于在比较有不同均值或不同量纲的两组数据的情况。例 1 中如甲、乙两组平均数不同，则应改用变异系数来比较甲、乙两人的技术水平。

例 3 某市 20 岁男子 100 人，其身高的均数为 166.06cm，标准差为 4.95cm；其体重均数为 53.72kg，标准差为 4.96kg。试比较身高与体重的变异程度是否认为相同。

解：由于单位不同，不能直接比较标准差，而应比较其变异系数。

身高：
$$CV=\frac{4.95}{166.06}\times 100\%=2.98\%$$

体重：
$$CV=\frac{4.96}{53.72}\times 100\%=9.23\%$$

可见体重的变异程度较大，即数据较离散，或者说身高比体重稳定。

三、偏态系数(Bias coefficient)

平均数和标准差揭示了频数分布的集中趋势和离中趋势。在对频数分布状况有了初步认识的基础上，还应研究频数分布的形态是对称分布还是偏态分布，即计算其偏度。偏度是指频数分布非对称的偏态方向程度。如前所述，偏态分布按其算术平均数与众数的大小关系，可分为右(正)偏或左(负)偏。因此，算术平均数与众数之间的距离可以作为测定频数分布偏态的尺度：

$$偏态=算术平均数(\bar{x})-众数(Mo)$$

如果 $\bar{x}>Mo$ 称此偏态为右(正)偏态；反之，当 $\bar{x}<Mo$ 称此偏态为左(负)偏态。$\bar{x}$ 与 Mo 的绝对差额越大，表明偏倚程度越大；$\bar{x}$ 与 Mo 的绝对差额越小，则表明偏倚程度越小。但由于偏态是以绝对数表示的，不能直接用于比较具有不同计量单位的次数分布的偏态，为此将偏态的值用标准差除之，即得其偏态系数或偏度，用 α 表示。其计算公式为

$$\alpha=\frac{\bar{x}-Mo}{\sigma} \tag{1.13}$$

α 取值一般应在 0～±3 之间。$\alpha=0$ 表示对称分布；$\alpha=3$ 与 $\alpha=-3$，分别表示极右偏态和极左偏态。