“统计学具有处理复杂问题的非凡能力，当科学的探索者在前进的过程中荆棘载途时，惟有统计学可以帮助他们打开一条通道。”

“很难理解为什么统计学家通常限制自己的调查于平均数，而不着迷于更广泛的考虑。对于变化的魅力，他们的灵魂看来如同平坦的英格兰国家之一的当地人的一样迟钝，那些当地人关于瑞士的回顾是，如果可以将它的山脉扔进它的湖泊，那么两种讨厌的东西将立即去除。”

——F. 高尔顿

第三章　变量分布特征的描述

本章介绍如何对变量分布的特征进行描述，内容包括集中趋势与平均指标、离中趋势与离散指标、分布形状与形状指标三大方面。本章内容对于以后各章的学习非常重要，具体要求：①理解变量分布三大特征即集中趋势、离中趋势和分布形状的的含义；②理解平均指标、离散指标和形状指标的意义与作用；③熟练掌握各种平均数的计算方法并加以正确的应用，科学理解加权平均数中权数的意义，正确认识算术平均数与调和平均数之间的应用关系，以及算术平均数、中位数和众数三者之间的数量关系；④熟练掌握各种离散指标的计算方法并加以正确的应用，尤其是要深刻理解方差、标准差和离散系数的内涵；⑤熟练掌握偏度系数和峰度系数的计算方法并加以正确的应用，尤其是要了解动差的含义。

第一节　集中趋势的描述

变量分布特征可以从以下三个方面加以描述：一是变量分布的集中趋势，反映变量分布中各变量值向中心值靠拢或聚集的程度；二是变量分布的离中趋势，反映变量分布中各变量值远离中心值的程度；三是变量分布的形状，反映变量分布的偏斜程度和尖陡程度。

一、集中趋势与平均指标

集中趋势亦称为趋中性，是指变量分布以某一数值为中心的倾向。作为中心的数值就称为中心值，它反映变量分布中心点的位置所在。对集中趋势的描述，就是要寻找变量分布的中心值或代表值，以反映某变量数值的一般水平。对于绝大多数统计变量来说，总是接近中心值的变量值居多，远离中心值的变量值较少，使得变量分布呈现出向中心值靠拢或聚集的态势，这种态势就是变量分布的集中趋势。

变量分布的集中趋势要用平均指标来反映。平均指标是将变量的各变量值差异抽象化、以反映变量值一般水平或平均水平的指标，也就是反映变量分布中心值或代表值的指标。平均指标的具体表现称为平均数，平均数因计算方法不同可分为数值平均数和位置平均数两类。数值平均数是指根据变量的所有数据计算而

得的平均数，主要有算术平均数、调和平均数和几何平均数等几种。位置平均数是指根据变量分布特征直接观察或根据变量数列部分处于特殊位置的变量值来确定的平均数，主要有中位数和众数等。

平均指标在统计研究中应用很广，其作用主要有以下几个方面：

（1）通过反映变量分布的一般水平，帮助人们对研究现象的一般数量特征有一个客观的认识。例如，要想了解某城市居民的收入水平，一一列出每家每户每人的收入显然是不可能、也不必要的，只要计算平均指标就可以了解该城市居民收入高低的基本状况。

（2）利用平均指标可以对不同空间的发展水平进行比较，消除因总体规模不同而不能直接比较的因素，以反映他们之间总体水平上存在的差距，进而分析产生差距的原因。

（3）利用平均指标可以对某一现象总体在不同时间上的发展水平进行比较，以说明这种现象发展变化的趋势或规律性。

（4）利用平均指标可以分析现象之间的依存关系或进行数量上的推算。例如将某城市样本居民按收入分组，计算出各组居民的平均收入与平均消费支出，就可以观察居民消费支出与收入之间的依存关系，还可以以样本居民的平均收入、平均消费支出去推算（估计）该城市居民的平均收入、平均消费支出。

（5）平均指标还可以作为研究和评价事物的一种数量标准或参考。在比较、评价不同总体的水平时，不能以各总体某一个体的水平为依据，而要看总体平均水平；在研究、评价个体事物在同类事物中的水平时，也必须以总体的平均水平为依据。在各项管理工作中，各种定额多是以实际平均数为基础来制定的。

二、数值平均数

（一）算术平均数

算术平均数也称为均值，是变量的所有取值的总和除以变量值个数的结果。算术平均数是统计中最为常用的用以描述集中趋势的平均数，因为它的计算方法客观上符合许多现象个体与总体之间存在的数量关系，即总体中每个个体标志值的算术和（即变量的各个变量值的算术和）等于总体标志总量（即变量值总和），把总体标志总量除以总个体数（即总体容量）就可以消除个体标志值之间的差异而体现出总体的一般水平。例如，某公司职工的工资总额是每个职工工资额的加总，职工的平均工资就等于职工工资总额除以公司职工人数。

由于掌握的资料不同，算术平均数可以分为简单算术平均数和加权算术平均数两种。

1.简单算术平均数

简单算术平均数是根据未分组数据计算的，即直接将变量的每个变量值相加，除以变量值的个数。若以 x 表示变量，以 x_i 表示第 i 个变量值（i=1，2，…，n），以 $\bar{x}$ 表示算术平均数，以 n 表示变量值个数，则简单算术平均数的计算公式为：

$$\bar{x}=\frac{x_1+x_2+\cdots+x_n}{n}=\frac{\sum_{i=1}^{n}x_i}{n} \quad （可简记为 \bar{x}=\frac{\sum x_i}{n}） \qquad (3\text{-}1)$$

【例 3-1】某高校学生男子篮球队 10 名队员的身高（单位：厘米）分别为 185，181，188，182，182，186，183，183，186，189，则该校学生男子篮球队队员的平均身高为：

$$\bar{x}=\frac{\sum x_i}{n}=\frac{185+181+188+182+182+186+183+183+186+189}{10}=184.5\text{（厘米）}$$

2.加权算术平均数

加权算术平均数是根据变量数列计算的，即以各组变量值（或组中值）乘以相应的频数求出各组标志总量，加总各组标志总量得出总体标志总量，再用总体标志总量除以总频数。若以 x_i 表示第 i 组的变量值（或组中值）（i=1，2，…，k），以 f_i 表示第 i 组的频数（i=1，2，…，k），以 k 表示分组数，则加权算术平均数的计算公式为：

$$\bar{x}=\frac{x_1f_1+x_2f_2+\cdots+x_kf_k}{f_1+f_2+\cdots+f_k}=\frac{\sum_{i=1}^{k}x_if_i}{\sum_{i=1}^{k}f_i}\quad\text{（可简记为 }\bar{x}=\frac{\sum x_if_i}{\sum f_i}\text{）}\qquad\text{（3-2）}$$

【例 3-2】某进出口公司 28 位业务员某年完成出口额的分组数据如表 3-1 所示，要求计算平均每人完成的年出口额。

表 3-1　　某进出口公司 28 位业务员完成出口额频数数列

年出口额（万元）	业务员人数
300	2
310	4
320	6
330	10
340	5
350	1
合计	28

根据表 3-1 数据可计算该进出口公司平均每人完成的年出口额为：

$$\bar{x}=\frac{\sum x_if_i}{\sum f_i}=\frac{300\times2+310\times4+320\times6+330\times10+340\times5+350\times1}{28}$$

=325.36（万元）

计算加权算术平均数时，有两个问题需要加以说明：

（1）关于权数问题。从公式可以看出，$\bar{x}$ 的大小不仅受变量值 x_i 大小的影响，而且受各组频数 f_i 大小的影响。不难发现，频数大的组的变量值对平均数的影响大，频数小的组的变量值对平均数的影响就小。当较大变量值出现的频数较大时，平均数就接近于变量值大的一端，而当较小变量值出现的频数较大时，平均数就接近于变量值小的一端。显然，各组频数对加权算术平均数的高低起着一种权衡轻重的作用，所以把 f_i 称为权数。可见，加权算术平均数是考虑了权数作用的算

术平均数。权数的选择必须考虑其与变量值之间的联系关系，即必须使$\sum x_i f_i$是计算算术平均数的真实的总体标志总量，符合实际意义。

加权算术平均数的权数除了用绝对数形式的频数f_i表示外，直接体现权数实质的是相对数形式的频率$f_i / \sum_{i=1}^{k} f_i$，即权数系数，因为相对数形式的权数有一个重要的性质，那就是各组权数之和等于1。因此，如果已知各组的频率，我们可以直接利用权数系数来求加权算术平均数，即加权算术平均数等于各组变量值与其权数系数乘积的总和：

$$\bar{x} = \frac{\sum_{i=1}^{k} x_i f_i}{\sum_{i=1}^{k} f_i} = \sum_{i=1}^{k} x_i \frac{f_i}{\sum_{i=1}^{k} f_i} \quad （可简记为 \bar{x} = \sum x_i \frac{f_i}{\sum f_i}） \qquad （3\text{-}3）$$

【例 3-3】根据例 3-2 的表 3-1 数据，可计算各组的频率如表 3－2 所示。

表 3-2　　某进出口公司 28 位业务员完成出口额频率数列

年出口额（万元）	业务员人数比重
300	0.0714
310	0.1429
320	0.2143
330	0.3571
340	0.1786
350	0.0357
合计	1.00000

根据表 3-2 数据可计算加权算术平均数为：

$$\bar{x} = \sum x_i \frac{f_i}{\sum f_i}$$

=300×0.0714+310×0.1429+320×0.2143+330×0.3571+340×0.1786+350×0.0357

=325.36（万元）

计算结果完全相同。

（2）关于按组距式数列计算加权算术平均数的问题。在组距式数列中，需要先计算各组的组中值作为各组的变量值，再按加权算术平均数的公式进行计算。应当指出的是，由于组中值是以假定各组的变量值均匀分布为前提的，因此利用组中值计算的加权算术平均数只是平均数的近似值。一般地，组距越小，计算结果越接近实际的平均数。

【例 3-4】根据表 2-3 数据计算某年年底某高校在职教师平均年龄。

根据表 2-3 数据可得平均年龄的计算表如表 3-3 所示。

表 3-3　　某年年底某高校在职教师平均年龄计算表

教师按年龄分组	组中值 x	人数（人） f	xf	频率 $\frac{f}{\sum f}$	$x\frac{f}{\sum f}$
30 岁以下	25	201	5025	0.1914	4.7850
30~40 岁	35	317	11095	0.3019	10.5665
40~50 岁	45	366	16470	0.3486	15.6870
50~60 岁	55	151	8305	0.1438	7.9090
60 岁以上	65	15	975	0.0143	0.9295
合计	—	1050	41870	1.0000	39.8800

平均年龄为：

$$\bar{x}=\frac{\sum x_i f_i}{\sum f_i}=\frac{41870}{1050}=39.88\text{（岁）}$$

或

$$\bar{x}=\sum x_i \frac{f_i}{\sum f_i}=39.88\text{（岁）}$$

这是一个近似结果。

3.算术平均数的数学性质

为了更好地理解和运用平均数，有必要了解算术平均数以下两条重要性质：

1.各变量值与算术平均数的离差之和等于零，即：

$$\sum(x_i-\bar{x})=0\text{（对于简单算术平均数）}\qquad(3\text{-}4)$$

或

$$\sum(x_i-\bar{x})f_i=0\text{（对于加权算术平均数）}\qquad(3\text{-}5)$$

2.各变量值与算术平均数的离差平方之和为最小值，即：

$$\sum(x_i-\bar{x})^2=\text{最小值}\qquad(3\text{-}6)$$

或

$$\sum(x_i-\bar{x})^2\le\sum(x_i-x_0)^2\qquad(3\text{-}7)$$

只有当 $\bar{x}=x_0$ 时，等号成立。

4.算术平均数的优缺点

算术平均数具有以下几个优点：一是可以利用算术平均数来推算总体标志总量，因为算术平均数与变量值个数之乘积等于总体标志总量（变量值总和）；二是由算术平均数的两个数学性质可知，算术平均数在数理上具有无偏性与有效性（方差最小性）的特点，这使得算术平均数在统计推断中得到了极为广泛的应用。三是算术平均数具有良好的代数运算功能，即分组算术平均数的算术平均数等于总体算术平均数。例如，某大学某年级某专业有两个班级，分别有 38 人和 42 人，某学期期末数学考试的算术平均成绩分别为 82 分和 85 分，则可计算该大学该年

级该专业某学期期末数学考试的总算术平均成绩为（38×82+42×85）÷80=83.575分。正因为如此，在实际中算术平均数比其他平均数得到更为广泛的应用。

但算术平均数也有其局限性，主要表现在以下两个方面：一是算术平均数易受特殊值（特大或特小值）的影响，当变量存在少数几个甚至一个特别大或特别小的变量值时，就会导致算术平均数迅速增大或迅速变小，从而影响对变量值一般水平的代表性。例如，某个体经营户户主的月收入为 30000 元，四位帮工的月收入分别为 1000 元、1000 元、1200 元和 1400 元，计算四位帮工的平均月收入为 1150 元，如果加上户主计算五位的平均月收入则为 6920 元，一下增加了 5770 元。很显然，6920 元这个平均数对于帮工和户主都不具有代表性，因为他们的实际月收入与该平均数的距离都非常大，原因就在于户主与帮工不具有同质性。所以，在计算算术平均数时如果遇到极端值，应该分析其原因，必要时（对于非同质的变量值）应该加以剔除。二是根据组距数列计算算术平均数时，由于组中值具有假定性而使得计算结果只是一个近似值，尤其是当组距数列存在开口组时，算术平均数的准确性会更差。

（二）调和平均数

调和平均数是平均数的一种。从数学形式上看，调和平均数具有独立的形式，它是变量值的倒数的算术平均数的倒数，也称为倒数平均数。但在实际应用中，它则是更多地作算术平均数的变形而存在。在计算平均数时，当我们不知道变量值个数（即总体总频数），而只知道各组变量值与各组标志总量（即各组变量总值）时，就要先以各组标志总量除以各组变量值求出各组频数；然后再以各组标志总量之和除以各组频数之和，这样所计算的平均数就叫做调和平均数。调和平均数也有简单调和平均数和加权调和平均数两种。

1.简单调和平均数

当各组的标志总量相等时，所计算的调和平均数称为简单调和平均数。设总体分为 k 个组，每个组的标志总量都为 m，则总体标志总量为 km。现仍以 x 表示各组变量值，以 H 表示调和平均数，则简单调和平均数的计算公式为：

$$H=\frac{km}{\frac{m}{x_1}+\frac{m}{x_2}+\cdots+\frac{m}{x_k}}=\frac{k}{\sum_{i=1}^{k}\frac{1}{x_i}} \quad （可简记为 H=\frac{k}{\sum\frac{1}{x_i}}） \qquad （3-8）$$

【例 3-5】市场上某种蔬菜的价格是早市每公斤 1.25 元，午市每公斤 1.20 元，晚市每公斤 1.10 元。若早、中、晚各买 10 元钱的蔬菜，问所购买蔬菜的平均价格是多少？

蔬菜的平均价格是总购买金额除以总购买数量。该例中有 3 个组，各组标志总量（购买金额）都为 10 元，各组变量值（蔬菜价值）分别为 1.25 元，1.20 元和 1.10 元，但不知道所购买蔬菜的数量，所以要先分别计算出各组的蔬菜购买数量，即 $\frac{10}{1.25}$、$\frac{10}{1.20}$ 和 $\frac{10}{1.10}$ 公斤，最后可计算出所购买蔬菜的平均价格为：

$$H=\frac{k}{\sum\frac{1}{x_i}}=\frac{30}{\frac{10}{1.25}+\frac{10}{1.20}+\frac{10}{1.10}}=\frac{30}{25.42}=1.180\text{（元/公斤）}$$

如果采用简单算术平均数计算，则所购买蔬菜的平均每公斤价格为：

$$\bar{x}=\frac{\sum x_i}{n}=\frac{1.25+1.20+1.10}{3}=1.183\text{（元/公斤）}$$

结果为什么不一样（虽然很接近）？因为本例实际上是花了 30 元钱购买了 25.42 公斤蔬菜，而不是花了 3.55 元买了 3 公斤蔬菜，所以简单算术平均数的结果 1.183 元/公斤是错误的。

2.加权调和平均数

当各组的标志总量不相等时，所计算的调和平均数要以各组的标志总量为权数，其结果即为加权调和平均数。若以 m_i 表示各组标志总量，则加权调和平均数的计算公式为：

$$H=\frac{m_1+m_2+\cdots+m_k}{\frac{m_1}{x_1}+\frac{m_2}{x_2}+\cdots+\frac{m_k}{x_k}}=\frac{\sum_{i=1}^{k}m_i}{\sum_{i=1}^{k}\frac{m_i}{x_i}}\quad\text{（可简记为 }H=\frac{\sum m_i}{\sum\frac{m_i}{x_i}}\text{）}\qquad(3\text{-}9)$$

【例 3-6】市场上某种蔬菜的价格是早市每公斤 1.25 元，午市每公斤 1.20 元，晚市每公斤 1.10 元。现若早、中、晚分别购买 15 元、12 元和 10 元钱的蔬菜，问所购买蔬菜的平均价格是多少？

与例 3-5 相比，早、中、晚购买蔬菜的金额不一样了，不再都是 10 元，此时平均价格会发生什么变化呢？不难计算，此时所购买蔬菜的平均价格为：

$$H=\frac{\sum m_i}{\sum\frac{m_i}{x_i}}=\frac{15+12+10}{\frac{15}{1.25}+\frac{12}{1.20}+\frac{10}{1.10}}=\frac{37}{31.09}=1.19\text{（元/公斤）}$$

计算结果显示，平均价格比例 3-5 上升了 0.01 元/公斤。为什么蔬菜价格未变，平均价格却上升了？原因就在于早、中、晚购买的金额不同，早市的价格最高且购买的金额最多，午市的价格次高且购买金额次多，晚市的价格最低且购买金额最少，所以与例 3-5 的简单调和平均数相比，平均价格就偏向于高的一端了。显然，购买金额就起到了权数的作用。更一般地说，加权调和平均数的权数作用是通过各组的标志总量 m 来体现的。

对于组距式数列，要先以各组的组中值作为各组的变量值 x，然后按照上述计算公式和步骤计算加权算术平均数。

加权调和平均数与加权算术平均数的区别就在于计算过程中应用数据条件的不同。前者以各组标志总量（$m_i=x_i f_i$）为权数，后者以各组频数（f_i）为权数。但他们都符合于总体标志总量与总体总频数的对比关系。事实上，两者是可以相互变通的，即：

$$\frac{\sum m_i}{\sum\frac{m_i}{x_i}}=\frac{\sum x_i f_i}{\sum\frac{x_i f_i}{x_i}}=\frac{\sum x_i f_i}{\sum f_i}\qquad(3\text{-}10)$$

所以对于同一现象，计算加权调和平均数与计算加权算术平均数的结果是相等的，无非是因数据条件不同而采用了不同的计算形式。

3.由相对数或平均数计算平均数

有时，我们需要根据相对数或平均数来计算平均数。例如，根据各零售分店的计划完成程度来计算全公司的计划完成程度；根据各企业的职工平均工资来计算全公司的职工平均工资等。这时总体平均数的计算要依所掌握的权数资料不同采取不同的方法。如果所掌握的权数资料是相对数或平均数的母项数值，要用加权算术平均数；如果所掌握的权数资料是相对数或平均数的子项数值，则要用加权调和平均数。需要强调的是，在以相对数或平均数计算平均数时，不论是用加权算术平均数公式还是用加权调和平均数公式，都要从相对数或平均数指标本身的经济含义出发来计算，这是一个很重要的原则。

（1）由相对数计算平均数

我们通过具体例子来加以说明。

【例 3-7】某市某商业零售公司所属的 20 家分店的销售计划完成情况及计划销售额如表 3-4 所示，要求计算全公司的平均计划完成程度。

表 3-4　　某市某零售公司 20 家分店销售计划完成情况

计划完成程度（%）	商店数（个）	计划销售额（万元）
80～90	2	100
90～100	3	105
100～110	8	480
110～120	4	260
120～130	3	200
合计		1145

计算计划完成程度的基本公式是：

$$\text{计划完成程度}=\frac{\text{实际完成数}}{\text{计划数}}\times 100\%$$

因此，在计算平均销售计划完成程度时不能以商店数为权数。由于我们所掌握的资料是相对数的母项数值即计划销售额，所以，应该以计划销售额为权数，采用加权算术平均数公式来计算销售计划平均完成程度。在计算出每组计划完成程度的组中值后，即可计算出全公司的平均计划完成程度。计算数据如表 3-5 所示。

表 3-5　　某市某零售公司平均销售计划完成程度计算数据

计划完成程度（%）	组中值（%） x_i	商店数（个）	计划销售额（万元） f_i	实际销售额（万元） $x_i f_i$
80～90	85	2	100	85.00
90～100	95	3	105	99.75
100～110	105	8	480	504.00
110～120	115	4	260	299.00
120～130	125	3	200	250.00
合计			1145	1237.75

全公司平均计划完成程度 $\bar{x}=\frac{\sum x_i f_i}{\sum f_i}=\frac{1237.75}{1145}\times 100\%=108.10\%$

如果只知道各组的实际销售额数据，而无计划销售额数据，那么我们所掌握的是计划完成程度相对数的子项数值，这时就应该以实际销售额为权数，采用加权调和平均数的公式来计算全公司的平均计划完成程度。原始数据及计算数据如表 3-6 所示。

表 3-6　　某市某零售公司各分店销售计划完成情况及计算数据

计划完成程度（%）	组中值（%） x_i	商店数（个）	实际销售额（万元） m_i	计划销售额（万元） $\frac{m_i}{x_i}$
80～90	85	2	85.00	100
90～100	95	3	99.75	105
100～110	105	8	504.00	480
110～120	115	4	299.00	260
120～130	125	3	250.00	200
合计			1237.75	1145

全公司平均计划完成程度 $H=\frac{\sum m_i}{\sum \frac{m_i}{x_i}}=\frac{1237.75}{1145}\times 100\%=108.10\%$

需要补充说明的是，全公司的平均计划完成程度实际上就是该公司总的计划完成程度，所以由相对数所计算的平均数实际上就是总的相对数。

（2）由平均数计算平均数

我们仍然通过具体例子来加以说明。

【例 3-8】某车间各班组工人的平均劳动生产率和实际工时数据如表 3-7 所示，要求计算车间平均劳动生产率。

表 3-7　　某车间各班组平均劳动生产率数据

班组	平均劳动生产率（件/工时）	实际工时（小时）
1	12	200
2	16	320
3	20	300
4	28	190
合计		1010

平均劳动生产率的计算公式为：

$$\text{平均劳动生产率}=\frac{\text{实际产品总量}}{\text{实际工时}}\times 100\%$$

由于我们掌握的资料是平均数的母项数值即实际工时数，因而应该以实际工

时数为权数，采用加权算术平均数的形式来计算平均劳动生产率。计算数据如表3-8所示。

表 3-8　　某车间平均劳动生产率计算数据

班组	平均劳动生产率（件/工时） x_i	实际工时 f_i	实际产品总量（件） $x_i f_i$
1	12	200	2400
2	16	320	5120
3	20	300	6000
4	28	190	5320
合计		1010	18840

车间平均劳动生产率 $\bar{x}=\frac{\sum x_i f_i}{\sum f_i}=\frac{18840}{1010}$=18.65（件/工时）

如果已知实际产品产量数据，而无实际工时数据，那么我们所掌握的是平均数的子项数值，这时就应该以实际产品产量为权数，采用加权调和平均数的形式来计算车间平均劳动生产率。原始数据和计算数据如表3-9所示。

表 3-9　　某车间各班组平均劳动生产率及计算数据

班组	平均劳动生产率（件/工时） x_i	实际产品总量（件） m_i	实际工时（小时） $\frac{m_i}{x_i}$
1	12	2400	200
2	16	5120	320
3	20	6000	300
4	28	5320	190
合计		18840	1010

车间平均劳动生产率 $H=\frac{\sum m_i}{\sum \frac{m_i}{x_i}}=\frac{18840}{1010}$=18.65（件/工时）

同样需要补充说明的是，车间的平均劳动生产率实际上就是该车间总的平均劳动生产率，所以由平均数所计算的平均数实际上就是总的平均数。

（三）几何平均数

几何平均数是计算平均比率或平均速度常用的一种方法，例如用于计算水平法的平均发展速度、流水作业生产的产品平均合格率、复利法的平均利率等。根据所掌握的数据条件不同，几何平均数也可以分为简单几何平均数和加权几何平均数两种。

1.简单几何平均数

简单几何平均数就是变量的n个变量值连乘积的n次方根。若以x_i表示变量的第i个变量值（i=1，2，3，…，n），以G表示几何平均数，则简单几何平均数的

计算公式为：

$$G=\sqrt[n]{x_1\cdot x_2\cdot x_3\cdot\cdots\cdot x_n}=\sqrt[n]{\prod_{i=1}^{n}x_i}\quad（可简记为 G=\sqrt[n]{\prod x_i}）\qquad（3\text{-}11）$$

【例 3-9】某机械厂五个流水作业车间的合格品率分别为 96％、94％、95％、95％和 96%，则五个车间合格品率的平均数（即全厂的平均生产合格率）为：

全厂平均合格品率 $G=\sqrt[5]{96\%\times94\%\times95\%\times95\%\times96\%}=95.20\%$

但要注意的是，该厂总的合格率为 $96\%\times94\%\times95\%\times95\%\times96\%=78.18\%$，两者相差甚大。

2.加权几何平均数

当计算几何平均数的各种变量值出现的次数不等，即数据经过了统计分组时，则应采用加权几何平均数。若以 x_i 表示第 i 组的变量值（i=1，2，…，k），以 f_i 表示第 i 组的频数（i=1，2，…，k），以 k 表示分组数，则加权几何平均数的计算公式为：

$$G=\sqrt[\sum_{i=1}^{k}f_i]{x_1^{f_1}\cdot x_2^{f_2}\cdot x_3^{f_3}\cdot\cdots\cdot x_k^{f_k}}=\sqrt[\sum_{i=1}^{k}f_i]{\prod_{i=1}^{k}x_i^{f_i}}\quad（可简记为 G=\sqrt[\sum_{i=1}^{k}f_i]{\prod x_i^{f_i}}）\qquad（3\text{-}12）$$

【例 3-10】某企业最近 10 年销售收入的年发展速度如表 3-10 所示，求年平均发展速度。

表 3-10　　某企业最近 10 年销售收入年发展速度数据

年发展速度（％）x_i	**105**	**106**	**107**	**108**	**109**
年数（频数）f_i	**3**	**3**	**2**	**1**	**1**

该企业最近 10 年销售收入的年平均发展速度为：

$$G=\sqrt[\sum_{i=1}^{k}f_i]{\prod x_i^{f_i}}=\sqrt[10]{105\%^3\times106\%^3\times107\%^2\times108\%\times109\%}=106.39\%$$

（四）算术平均数、调和平均数和几何平均数的数学关系

从数学上看，算术平均数、调和平均数和几何平均数都是幂平均数的一种。幂平均数的定义是：

$$\overline{x^t}=\sqrt[t]{\frac{\sum x^t}{n}}\qquad（3\text{-}13）$$

当 t=1 时，幂平均数就是算术平均数；当 t=−1 时，幂平均数就是调和平均数；当 t 趋向于 0 时，幂平均数的极限形式就是几何平均数。

由于幂平均函数是单调递增函数，所以 t 值越大幂平均数就越大，因此单从数学意义上看，算术平均数、调和平均数和几何平均数三者的大小关系是：

$$H\le G\le\overline{x}\qquad（3\text{-}14）$$

但在实际应用中这样的比较往往没有意义，因为对于任何一个计算对象一般都只适合采用一种方法来计算平均数，也就是说不同的平均数计算方法适合于不

同的计算条件，必须加以正确的选择。

三、位置平均数

（一）中位数与分位数

1.中位数

中位数是变量的所有变量值按定序尺度排序后，处于中间位置的变量值。由于它居于数列的中间位置，所以在某些情况下可以用来代表变量值的一般水平。中位数既可用以测定定量变量的集中趋势，也可用以测定定序变量的集中趋势，但不适用于定类变量。

中位数的确定，因所掌握的数据条件不同而分为两种情况：一是根据变量未经分组的原始数据来确定；二是根据变量分布数列来确定。

（1）根据未经分组的原始数据来确定

在变量数据未经分组的情况下，先将变量的n个数据按大小、强弱等顺序排列，确定中位数的位置$\frac{n+1}{2}$，然后确定中位数。

假设变量的n个数据按大小、强弱等顺序排列后的结果为：$x_{(1)}$，$x_{(2)}$，$x_{(3)}$，…，$x_{(n)}$，以m_e表示中位数，则

$$m_e=\begin{cases} x_{(\frac{n+1}{2})}, & n\text{为奇数} \\ \frac{1}{2}\left\{x_{(\frac{n}{2})}+x_{(\frac{n}{2}+1)}\right\}, & n\text{为偶数} \end{cases} \qquad (3\text{-}15)$$

【例3-11】7名体育竞技专家对某运动员协调性的评级依次为：B，A^-，A，A，A，A^+，A^+，问该运动员协调性评级的中位数是多少？

本例中，n=7，中位数位置是4，所以中位数是m_e=A 。

【例3-12】根据例3-1数据确定该校学生男子篮球队员身高的中位数。

本例中，n=10，中位数位置为5.5，所以中位数是身高排序后第5、第6两名队员身高的平均数。

10名队员的身高（单位：厘米）由低到高排序为：181，182，182，183，183，185，186，186，188，189。第5、第6两名队员的身高分别为183和185厘米，所以该校学生男子篮球队员身高的中位数是：

$$m_e=\frac{183+185}{2}=184\text{（厘米）}$$

（2）根据变量分布数列确定中位数

在单项式数列中，先按（$\sum f_i$+1）/2 来确定中位数位置，然后对数列中的各组频数进行向上累计或向下累计，当某一组的累计频数大于或等于（$\sum f_i$+1）/2

时，该组的变量值就是中位数。

【例 3-13】某车间 150 名工人的日装配量如表 3-11 所示，要求确定工人日装配量的中位数。

表 3-11　　某车间 150 名工人日装配量及累计频数

日装配量（件）	工人数（频数）	向上累计频数	向下累计频数
22	10	10	150
23	10	20	140
24	40	60	130
25	50	110	90
26	30	140	40
27	10	150	10
合计	150		

根据所给数据可以计算中位数位置$=\frac{\sum f_i+1}{2}=\frac{150+1}{2}=75.5$。在表 3-11 中对各组频数进行向上累计或向下累计，向上累计至第四组（累计频数 110）或向下累计至第三组（累计频数 90），累计频数大于 75.5，所以工人日装配量的中位数就是 m_e=25（件）。

按组距数列计算中位数，首先要计算各组的累计频数，并按$\frac{\sum f_i}{2}$确定中位数的位置。然后找出中位数所在的位置，即累计次数大于或等于$\sum f/2$的组。最后，再用插值法按比例计算中位数的近似值。具体计算有下限和上限公式两种，结果是一样的。

中位数公式示意图如图 3-1 所示。

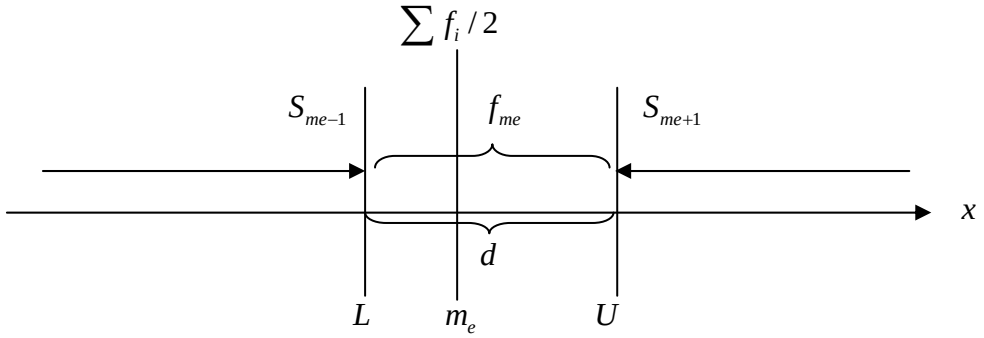


图 3-1　　中位数公式示意图

下限公式为：

$$m_e=L+\frac{\frac{\sum f_i}{2}-S_{me-1}}{f_{me}}\times d \qquad (3\text{-}16)$$

式中L为中位数所在组的下限，f_{me}为中位数所在组的频数，S_{me-1}为向上累计至中位数所在组下一组止的累计频数，d为中位数所在组的组距。

上限公式为：

$$m_e = U - \frac{\frac{\sum f_i}{2} - S_{me+1}}{f_{me}} \times d \qquad (3-17)$$

式中U为中位数所在组的上限，S_{me+1}为向下累计至中位数所在组上一组的累计频数。

【例 3-14】根据表 2-3 数据计算某年年底某高校在职教师年龄的中位数。

由表中数据可以计算中位数位置为$\sum f_i/2$＝1050/2＝525。根据表 2-4 可知，向上累计至第 3 组的累计频数（884）或向下累计至第 3 组的累计数（532）大于525，因而中位数所在组为 40～50 岁这一组，L=40，U=50，d=10。

由下限公式得：

$$m_e = 40 + \frac{\frac{1050}{2} - 518}{366} \times 10 = 40.19 \text{（岁）}$$

由上限公式得：

$$m_e = 50 - \frac{\frac{1050}{2} - 166}{366} \times 10 = 40.19 \text{（岁）}$$

（3）中位数的应用特点

中位数将按顺序排列的变量值分为了两部分，使得至少一半数值不比它大，至少一半数值不比它小。

中位数具有以下一些优点：一是中位数作为一种位置平均数，概念较为清晰，只要排列数据顺序，就可比较容易地加以确定；二是中位数不受变量数列中特殊值的影响，遇有特大值或特小值时，用中位数来表示现象的一般水平更具有代表性；三是组距数列出现开口组时，对中位数无影响；四是当某些变量不能表现为数值但可以定序时，不能计算数值平均数而可以确定中位数。

当然中位数也有局限性，一是中位数不能如算术平均数那样可以进行代数运算；二是除了变量数列的中间部分数值外，其他数值的变化都不对中位数产生影响，因此中位数的灵敏度较低。

2.分位数

分位数是将变量的数值按大小顺序排列并等分为若干部分后，处于等分点位置的数值。常用的分位数有四分位数、十分位数和百分位数，他们分别是将数值序列 4 等分、10 等分和 100 等分的 3 个点、9 个点和 99 个点上的数值。其中四分位数第 2 点的数值、十分位数第 5 个点的数值和百分位数第 50 个点的数值，就是中位数。所以，中位数就是一个特殊的分位数。

以四分位数为例，设Q_L，Q_M和Q_U分别表示第一个、第二个和第三个四分位

数，则他们的位置分别为：$\frac{n+1}{4}$，$\frac{2(n+1)}{4}$ 和 $\frac{3(n+1)}{4}$，根据位置即可确定各个四分位数。

【例 3-15】根据例 3-11 确定运动员协调性评级的第一个和第三个四分位数。

由于 n =7，所以第一个和第三个四分位数的位置分别是 2 和 6，由此可以确定第一个和第三个四分位数分别为 $Q_L = \mathrm{A}^-$ 和 $Q_U = \mathrm{A}^+$。

【例 3-16】根据例 3-1 数据确定该校学生男子篮球队员身高的第一个和第三个四分位数。

由于 n =10，所以第一个和第三个四分位数的位置分别是 2.75 和 8.25，由此可以确定：

第一个四分位数为 Q_L =182×0.75+182×0.25=182（厘米）

第三个四分位数为 Q_U =186×0.25+188×0.75=187.5（厘米）

同理，也可根据单项式数列和组距式数列确定第一个和第三个四分位数。例如，根据表 3-11 可以确定工人日装配量的第一个和第三个四分位数分别为 Q_L =24（件）和 Q_U =26（件）。根据表 2-3 和表 2-4，参照中位数公式可以确定某年年底某高校在职教师年龄的第一个和第三个四分位数分别为 Q_L =31.94（岁）和 Q_U =47.36（岁）。请读者自己加以验证。

确定各个四分位数后可以绘制如第二章所介绍的箱形图，当 n 为偶数时可据以观察变量分布中间 $\frac{n}{2}$ 或 $\frac{n}{2}+2$ 个变量值（不含第一、第三分位数本身，下同）的分布范围、中心位置和对称程度，当 n 为奇数时可据以观察变量分布中间 $\frac{n-1}{2}$ 或 $\frac{n+1}{2}$ 个变量值的分布范围、中心位置和对称程度。例如，例 3-15 的结果表明处于 A^- 与 A^+ 之间的变量值有 3 个，例 3-16 的结果表明处于 182 厘米与 187.5 厘米之间的变量值有 6 个（注意：数值序列中第二、第三个数值都是 182，但前者处于第一个四分位数以下，后者处于第一个、第三个四分位数之间）。当 n 或 $\sum f$ 很大时，我们可以说数值序列或变量数列中间约 50%的变量值在 Q_L 与 Q_U 之间，例如工人日装配量的例子中，我们可以说日装配量居中的约 50%工人的日装配量在 24~26 件之间（同样要注意：日装配量 24 件有 40 人，其中 17 人在第一个四分位数以下，23 人在第一个四分位数以上；日装配量 26 件有 30 人，其中 3 人在第三个四分位数以下，27 人在第三个四分位数以上，日装配量处于第一个、第三个四分位数之间的工人有 76 人）；某高校在职教师年龄的例子中，我们可以说年龄居中的约 50%教师的年龄在 31.94~47.36 岁之间。

（二）众数

众数是变量数列中出现次数最多、频率最高的变量值。在某些场合，众数可以用来反映现象的一般水平。例如城市居民家庭中，三口之家所占的比重明显高于其他家庭，因此 3 人就是城市居民家庭人数的众数，可以用它来表示城市居民家庭人数的一般水平。众数通常用 m_o 来表示。

众数可用以测定任何种类变量的集中趋势，包括定类变量和定序变量。例如，某班级要搞一次暑期社会实践活动，有 A、B、C、D、E 五种备选方案，经同学投票 B 方案得票明显高于其他方案，则 B 方案就是众数。再如，根据表 2-2 某年年底某高校在职教师职称分布数列可以看出，副教授职称的人数最多（382 人，占 36.38%），所以职称的众数就是副教授。

众数的确定方法因所掌握的数据条件不同而有所不同。根据单项式数列确定众数比较容易，只要找出出现频数最多或出现频率最高的变量值即可。例如，根据表 3-11 数据可以确定工人日装配量的众数是 25 件。

如果根据组距式数列来确定众数，则先要找出频数最多的一组作为众数组，然后运用公式来确定众数。

众数公式示意图如图 3-2 所示。

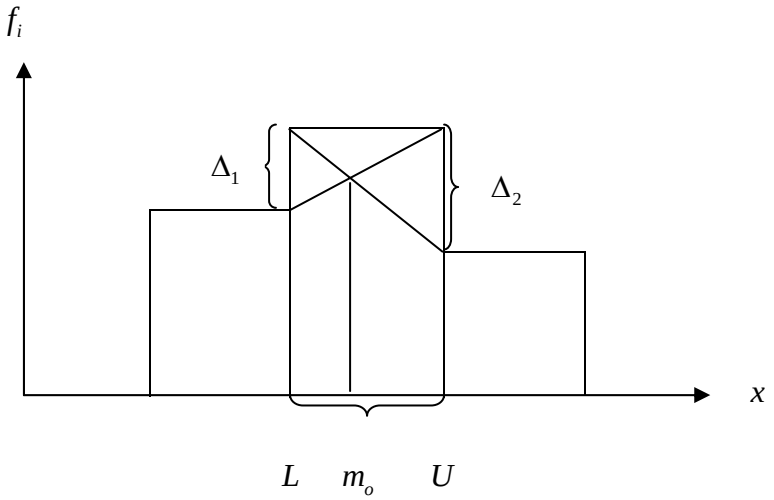


图 3-2　众数公式示意图

下限公式为：

$$m_0 = L + \frac{\Delta_1}{\Delta_1 + \Delta_2} \times d \qquad (3\text{-}18)$$

式中 Δ_1 为众数组频数与下一组频数之差，Δ_2 为众数组频数与上一组频数之差；L、d 的含义与中位数公式的相同。

上限公式：

$$m_0 = U - \frac{\Delta_1}{\Delta_1 + \Delta_2} \times d \qquad (3\text{-}19)$$

式中 U 的含义与中位数公式的相同。

【例 3-17】根据表 2-3 数据计算某年年底某高校在职教师年龄的众数。

根据表中数据可知：众数组为 40~50 岁这一组。L ＝40，U ＝50，Δ_1＝49，Δ_2＝215，d ＝10。

由下限公式得：

$$m_0 = L + \frac{\Delta_1}{\Delta_1 + \Delta_2} \times d = 40 + \frac{49}{49 + 215} \times 10 = 41.86\text{（岁）}$$

由上限公式得：

$$m_0 = U - \frac{\Delta_1}{\Delta_1 + \Delta_2} \times d = 50 - \frac{215}{49 + 215} \times 10 = 41.86\text{（岁）}$$

众数具有以下一些特点：一是众数也不受变量数列中特殊值的影响，用它来表示某些现象的一般水平会有较好的代表性；二是众数具有较广的应用面，可用于测定任何变量的集中趋势；三是众数只有在总频数充分多且某一组的频数明显高于其他组时才有意义，若各组的频数相差不多，则不能确定众数；四是有时一个变量数列会有两个组的频数明显最多，这就会有两个众数，该数列属于双众数数列。例如，英语专业与非英语专业的大学二年级学生参加同一英语水平测试，就可能出现双众数现象；再如现在一些高校招生，有的专业在第一批录取，有的专业在第二批次录取，那么全校新生的成绩分布也可能是双众数分布。五是众数也不能象算术平均数那样进行代数运算。

（三）中位数、众数和算术平均数的关系

中位数、众数、算术平均数三者在不同条件下均可代表变量的平均水平，均可用以反映变量分布的集中趋势。如果把三者结合起来，通过比较他们之间的数量关系，可以帮助我们更好地认识变量分布的特征。

1.在变量分布完全对称（即正态分布）时，中位数、众数和算术平均数三者完全相等，即 $\bar{x} = m_e = m_0$，如图 3-3 所示。

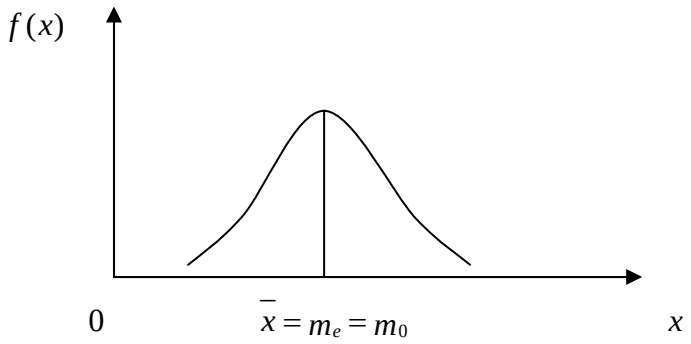


图 3-3　正态分布时中位数、众数和算术平均数的关系

2.在变量分布不对称（即偏态分布）时，中位数、众数和算术平均数三者之间存在着差异。当算术平均数受极大标志值一端的影响较大时，变量分布向右偏，三者之间的关系为：$m_0 < m_e < \bar{x}$，如图 3-4 所示。当算术平均数受极小标志值一端的影响较大时，变量分布向左偏，三者之间的关系为：$\bar{x} < m_e < m_0$，如图 3-5 所示。

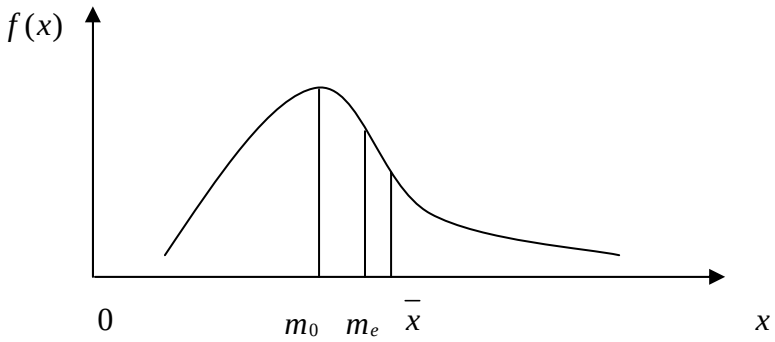


图 3-4　右偏分布时中位数、众数和算术平均数的关系

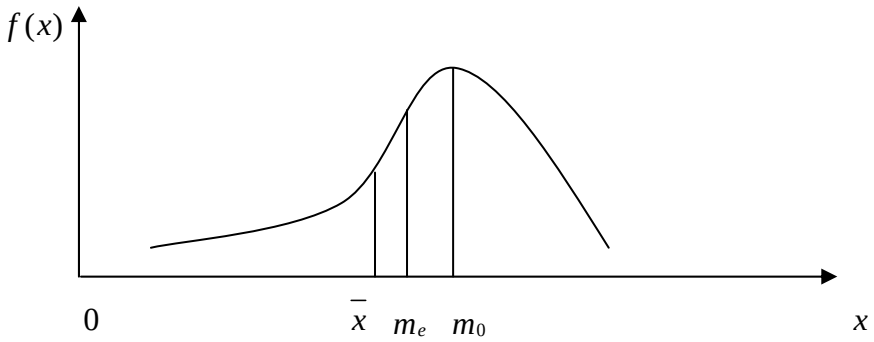


图 3-5　左偏分布时中位数、众数和算术平均数的关系

3.根据经验，在轻微偏态时，不论是左偏还是右偏，众数与算术平均数的距离约等于中位数与算术平均数的 3 倍，即：

$$m_0 - \bar{x} = 3(m_e - \bar{x}) \qquad (3\text{-}20)$$

利用这个公式，我们可以从已知的两个平均数来推算另一个平均数。

第二节　离中趋势的描述

一、离中趋势和离散指标

变量分布既有集中趋势的一面，又有离中趋势的一面。所谓离中趋势，就是变量分布中各变量值背离中心值的倾向。如果说集中趋势是总体或变量分布同质性的体现，那么离中趋势就是总体或变量分布变异性的体现。对离中趋势的描述，就是要反映变量分布中各变量值远离中心值或代表值的状况，以更客观地反映变量分布的特征。

变量分布的离中趋势要用离散指标来反映。离散指标就是反映变量值变动范围和差异程度的指标，即反映变量分布中各变量值远离中心值或代表值程度的指标，亦称为变异指标或标志变动度指标。常用的离散指标主要有：全距（亦称极差）、四分位差、异众比率、平均差、标准差、离散系数等。

利用离散指标，不仅可以看出变量分布的离中程度，而且与平均指标结合运

用，可以更正确地认识总体现象或变量分布的数量特征，对于科学管理与决策具有重要的意义。具体来说，离散指标的作用主要有以下几点：

（1）可以用来衡量和比较平均数的代表性

平均数掩盖了各变量值之间的差异，具有抽象性与代表性。平均数代表性的高低不是取决于它自己本身，而是取决于它背后各变量值之间的差异程度。如果变量的变动幅度大或各变量值之间的差异程度大，则平均数的代表性就小；如果变量的变动幅度小或各变量值之间的差异程度小，，则平均数的代表性就大。

（2）可以用来反映各种现象活动过程的均衡性、节奏性或稳定性

现象的活动过程通常都以平均数为中心而呈现出波动，波动的大小说明现象活动过程的均衡性、节奏性或稳定性的高低，而这种波动同样可以通过离散指标来反映。例如，国民经济发展过程中增长速度是否大起大落？股票价格的变化是否暴涨暴跌？计划执行过程是否忽松忽紧？产品生产质量是否稳定均匀？等等，都可利用离散指标来加以反映。

（3）为统计推断提供依据

在统计推断中，无论是抽样估计还是假设检验，离散指标都是必不可少的要素，也是得出统计推断结论或判断推断效果（例如估计效果、预测效果）的重要依据。

二、离散指标的测度

（一）全距

全距就是变量的最大值（$x_{\max}$）与最小值（$x_{\min}$）之差，也叫极差，表明变量的最大变动范围或绝对幅度。全距通常用R表示，即：

$$R = x_{\max} - x_{\min} \qquad (3\text{-}21)$$

全距一般只根据未分组数据或单项式数列计算。例如，根据例 3-1 可计算该高校学生男子篮球队员身高的全距为 8 厘米，根据例 3-13 可计算工人日装配量的全距为 5 件。对于组距式数列，全距只能根据最高组的上限减去最低组的下限来近似计算。

全距是测定变量分布离中趋势最简单的方法，在实际中也有众多的应用，例如每天天气预报中最高温与最低温之间的温差，股票市场中各种股票每天最高成交价与最低成交价之间的价差，人体血压中收缩压与舒张压之间的压差等，都是全距的表现。但由于全距只考虑了两个极端变量值之间的差距，没有利用全部变量值的信息，没有考虑变量中间分布的情况，所以不能充分反映全部变量值之间的实际差异程度，因而在应用上有一定的局限性。

由于全距只随两个极端值的变化而变化，缺乏稳定性，所以有时我们可以把变量的最高 5%（或 10%）数值的平均数与最低 5%（或 10%）数值的平均数之差作为全距。

（二）四分位差

四分位差是四分位数中第一个四分位数与第三个四分位数之差，也称为内距或四分间距，通常用Q_d表示，即：

$$Q_d = Q_U - Q_L \qquad (3\text{-}22)$$

例如，根据例 3-15 的结果可以计算运动员协调性评级的四分位差为

$Q_d = A^+ - A^-$=2 个等级；根据例 3-16 可计算该高校学生男子篮球队员身高的四分位差为 Q_d=187.5−182=5.5（厘米），根据上述某年年底某高校在职教师年龄分布的有关例子可计算教师年龄的四分位差为 Q_d=47.36−31.94=15.42（岁）。

四分位差通常与中位数相结合，用以表明变量分布中间 50%数值的离散程度，其值越小（越大），表明变量中间数值的分布越集中（越离散），中位数的代表性越好（越差）。

（三）异众比率

异众比率是分布数列中非众数组的频数与总频数之比，通常用 V_r 来表示，即：

$$V_r = \frac{\sum f_i - f_{mo}}{\sum f_i} = 1 - \frac{f_{mo}}{\sum f_i} \qquad (3\text{-}23)$$

其中 f_{mo} 为众数组的频数。

例如，根据例 3-13 的表 3-11 数据可以计算工人日装配量的异众比率为 V_r=0.67，根据表 2-13 数据可计算某年年底某高校在职教师年龄的异众比率为 V_r=0.65。

异众比率通常与众数相结合，用以表明众数代表性的高低。异众比率越大（越小），说明数列的分布越分散（越集中），众数的代表性就越差（越好）

（四）平均差

平均差是变量的各变量值与算术平均数离差绝对值的算术平均数，表明各变量值与算术平均数的平均差距，通常用 $A.D$ 来表示，即：

$$A.D = \frac{\sum_{i=1}^{n} |x_i - \bar{x}|}{n} \quad （根据未分组数据，可简记为 $A.D = \frac{\sum |x_i - \bar{x}|}{n}$） \qquad (3\text{-}24)$$

或

$$A.D = \frac{\sum_{i=1}^{k} |x_i - \bar{x}| f_i}{\sum_{i=1}^{k} f_i} \quad （根据变量数列，可简记为 $A.D = \frac{\sum |x_i - \bar{x}| f_i}{\sum f_i}$） \qquad (3\text{-}25)$$

平均差由于利用了全部数据信息，因而比全距、四分位差等更能比较客观反映变量分布的离散程度。平均差愈大，表示变量分布离散程度愈大；平均差愈小，则变量分布离散程度愈小。但由于对每一个离差都取了绝对值，因而数学处理不是很方便，数学性质也不是最优，应用上受到了一些限制。

【例 3-18】某企业工人日产量分组数据如表 3-12 所示，要求计算平均差。

表 3-12　　某企业工人日产量分组数据

工人日产量分组（件）	工人数
40 以下	10
40～50	20
50～60	15
60 以上	5
合计	50

根据表 3-12 可得到平均差计算表如表 3-13 所示。

表 3-13　　工人日产量平均差计算表

工人日产量分组（件）	组中值 x_i	工人数 f_i	$x_i f_i$	$\lvert x_i-\bar{x}\rvert$	$\lvert x_i-\bar{x}\rvert f_i$
40 以下	35	10	350	13	130
40～50	45	20	900	3	60
50～60	55	15	825	7	105
60 以上	65	5	325	17	85
合计		50	2400		380

根据表 3-13 可得：

$$\bar{x}=\frac{\sum x_i f_i}{\sum f_i}=\frac{2400}{50}=48\text{（件）}$$

$$A.D=\frac{\sum\left|x_i-\bar{x}\right| f_i}{\sum f_i}=\frac{380}{50}=7.6\text{（件）}$$

这说明平均每名工人的日产量与平均产量的差额为 7.6 件。

（五）方差和标准差

方差是变量的各变量值与其均值的离差平方的算术平均数，标准差则是方差的平方根。方差和标准差是测度变量分布离散程度最重要的指标，在统计学中具有非常重要的作用。方差通常 s^2 来表示，标准差则用 s 来表示。

方差的计算公式为：

$$s^2=\frac{\sum_{i=1}^{n}(x_i-\bar{x})^2}{n}\quad\text{（根据未分组数据，可简记为 }s^2=\frac{\sum(x_i-\bar{x})^2}{n}\text{）}\qquad(3\text{-}26)$$

或

$$s^2=\frac{\sum_{i=1}^{k}(x_i-\bar{x})^2 f_i}{\sum_{i=1}^{k} f_i}\quad\text{（根据变量数列，可简记为 }s^2=\frac{\sum(x_i-\bar{x})^2 f}{\sum f_i}\text{）}\qquad(3\text{-}27)$$

标准差的计算公式为：

$$s=\sqrt{\frac{\sum_{i=1}^{n}(x_i-\overline{x})^2}{n}}$$ （根据未分组数据，可简记为 $s=\sqrt{\frac{\sum(x_i-\overline{x})^2}{n}}$ ） （3-28）

或

$$s=\sqrt{\frac{\sum_{i=1}^{k}(x_i-\overline{x})^2 f_i}{\sum_{i=1}^{k} f_i}}$$ （根据变量数列，可简记为 $s=\sqrt{\frac{\sum(x_i-\overline{x})^2 f}{\sum f_i}}$ ） （3-29）

方差和标准差利用了全部数据信息，因而能准确反映变量分布的离散程度。方差或标准差愈大，表示变量分布离散程度愈大；方差或标准差愈小，则变量分布离散程度愈小。尤其是标准差与平均差相比，不仅具有平均差的优点，而且弥补了平均差的不足，再加上标准差的计量单位与变量相同，意义比方差明确，所以标准差在实践中得到了广泛的应用。

【例 3-19】根据例 3-18 的数据计算某企业工人日产量的方差和标准差。

根据表 3-12、表 3-13 和例 3-18 的计算结果，可得到方差和标准差的计算表如表 3-14 所示。

表 3-14　　工人日产量方差和标准差计算表

工人日产量分组（件）	组中值 x_i	工人数 f_i	离差 $(x_i-\overline{x})$	离差平方 $(x_i-\overline{x})^2$	离差平方×次数 $(x_i-\overline{x})^2 f_i$
40 以下	35	10	−13	169	1690
40～50	45	20	−3	9	180
50～60	55	15	7	49	735
60 以上	65	5	17	289	85
合计		50			2690

根据表 3-14 可得：

$$s^2=\frac{\sum_{i=1}^{k}(x_i-\overline{x})^2 f_i}{\sum_{i=1}^{k} f_i}=\frac{2690}{50}=53.8$$

$$s=\sqrt{\frac{\sum_{i=1}^{k}(x_i-\overline{x})^2 f_i}{\sum_{i=1}^{k} f_i}}=\sqrt{53.8}=7.33\text{（件）}$$

计算结果表明，工人日产量的标准差为 7.33 件。

关于方差和标准差还有以下两点需要说明：一是根据组距式数列计算的方差和标准差只是一个近似值，因为组中值成为了组内所有变量值的代表值，并没有

把组内各变量值的差异反映出来；二是在根据样本数据（甚至是有限总体数据）计算方差和标准差时，分母应该是$n-1$（$n=\sum f_i$），即样本方差和标准差的自由度为$n-1$，但当n很大时，可以忽略n与$n-1$之间的区别。

方差和标准差具有以下一些性质：

（1）常数的方差为 0。假设常数为a，常数的方差为s_a^2，则

$$s_a^2=0 \qquad (3\text{-}30)$$

（2）若$y=a+bx$，a,b为常数，则y的方差s_y^2与x的方差s_x^2之间的关系为：

$$s_y^2=b^2s_x^2。 \qquad (3\text{-}31)$$

（3）标准差s是计算标准化值的依据。假设变量的标准化统计量用Z表示，标准化值用Z_i表示，则

$$Z_i=\frac{x_i-\overline{x}}{s} \qquad (3\text{-}32)$$

Z服从均值为 0、标准差为 1 的标准正态分布，是无量纲。Z_i也叫标准得分或标准统计值。通过计算标准化值可以使处于不同均值水平、不同计量单位的变量值之间的比较成为了可能，使比较的对象找到同一标准的相对位置。例如，某班级某次期末数学考试的均值和标准差分别为 85 分和 7 分，英语考试的均值和标准差分别为 80 分和 6 分。周同学数学与英语的考试成绩分别为 92 分和 87 分，问周同学哪一门课程的成绩在班级中更好一些？可以计算周同学数学的标准化成绩为 1，英语的标准化成绩为 1.17，两个标准化成绩都大于 0，说明周同学两门课程的成绩都高于全班平均水平，并且英语成绩相对更好一些。

一般地，若变量分布呈对称的钟型分布，则分别有约 68%、95%和 99%的变量值处于以均值为中心加减 1 个标准差、2 个标准差和 3 个标准差的范围内，这是一个经验法则。如果某个变量值处于以均值为中心加减 3 个标准差的范围之外，即标准化值小于−3 或大于 3，那么这个变量值就称为离群值，属于测量、记录、输入有误，或来自非同类总体，或代表稀有事件。

（六）离散系数

全距、四分位差、平均差和标准差等都是反映变量分布离散程度的绝对指标，其数值大小取决于变量值本身水平即均值水平的高低，并且都有明确的计量单位。因此，不同均值水平和不同计量单位的绝对离散指标是不能直接比较的。为了不同变量分布之间离散程度的可比性，就必须消除不同均值水平和不同计量单位的影响，就应该计算相对离散指标。

相对离散指标也叫离散系数变异系数或标准差系数，是变量的标准差与均值之比，通常用V_s来表示，即：

$$V_s=\frac{s}{\overline{x}} \qquad (3\text{-}33)$$

离散系数越大，说明变量分布的离散程度越强，平均数的代表性越差；离散

系数越小，说明变量分布的离散程度越弱，平均数的代表性越好。

【例 3-20】根据例 3-18、例 3-19 的有关结果计算某企业工人日产量的离散系数。

根据例 3-18、例 3-19 的有关结果可计算离散系数为：

$$V_s = \frac{s}{\bar{x}} = \frac{7.33}{48} = 15.27\%$$

【例 3-21】如果又已知另一企业工人日产量的均值为 60 件，标准差为 8.6 件，问哪一个企业工人日产量的均值更具有代表性（或工人日产量更均匀）？

可以计算另一企业工人日产量的离散系数为$V_s = \frac{s}{\bar{x}} = \frac{8.6}{60} = 14.33\%$。不难发现，虽然另一企业的标准差（8.6 件）大于某企业（7.33 件），但离散系数却是另一企业（14.33%）小于某企业（15.27%），因此另一个企业工人日产量的均值更具有代表性（工人日产量更均匀）。

第三节　分布形状的描述

一、分布形状和形状指标

如前所述，变量分布的形状是各种各样的，有 J 型的、U 型的和钟型的等。仅就钟型分布而言，有的左右两侧完全对称，有的左偏，有的右偏。有的比较偏平，有的比较适中，有的则比较尖陡。分布形状不同，表明变量分布的内在结构也不同。为了全面了解变量分布的特征，我们不仅要观察其集中趋势和离中趋势，也要观察其形状。

变量分布的形状要用形状指标来反映。形状指标就是反映变量分布具体形状，即左右是否对称、偏斜程度与陡峭程度如何的指标。具体来说，变量分布的形状一般从对称性和陡峭性两方面来反映，因此形状指标也有两个方面：一是反映变量分布偏斜程度的指标，称为偏度系数；二是反映变量分布陡峭程度的指标，称为峰度系数。

偏度系数可以告诉我们变量分布是左偏还是右偏，即受低端变量值的影响大还是受高端变量值的影响大。而峰度系数则可以告诉我们分布是尖陡还是扁平，即频数（频率）分布绝大部分集中于众数附近还是各变值的频数（频率）相差不大（如果各变量值的频数或频率相等，则分布呈一条直线，无峰顶可言）。由此可见，形状指标与平均指标、离散指标一样，都是变量分布特征的重要体现。

二、偏度系数

偏度的概念首先由统计学家皮尔逊（Pearson）于 1895 年提出，是对变量分布对称性的测度，是指变量分布偏斜的方向及其程度。在本章第一节论述算术平均数、中位数和众数三者的关系时，曾经涉及到这个问题。图 3-1 表示变量分布对称无偏，图 3-2 表示变量分布向右偏斜（即右偏或正偏），图 3-3 表示变量分布向左偏斜（即左偏或负偏）。

偏度的测定是通过计算偏度系数来实现的，通常用S_k来表示。偏度系数的计算主要有以下三种方法：

（一）利用算术平均数与众数或中位数的离差求偏度系数

前面已提到，如果算术平均数、众数与中位数三者相等，则变量分布无偏；如果三者不相等，则变量分布有偏，而且三者之间的差距越大变量分布的偏度也越大。因此，我们可以利用算术平均数与众数或中位数的离差求偏度系数并标记为$S_k^{(1)}$，计算公式为：

$$S_k^{(1)}=\frac{\bar{x}-m_0}{s} \tag{3-34}$$

将$\bar{x}-m_o$除以标准差s，一是为了消除了不同计量单位的影响，二是为了把不可直接比较的绝对数转化为可相互比较的相对数。

一般情况下，偏度系数$S_k^{(1)}$的变动范围为（－3，3）。当$\bar{x}>m_o$时，$S_k^{(1)}$为正值，变量分布属于正偏；当$\bar{x}<m_o$时，$S_k^{(1)}$为负值，变量分布属于负偏；当$\bar{x}=m_o$时，$S_k^{(1)}$为 0，变量分布属于无偏（即对称分布）。$S_k^{(1)}$的绝对值越接近于 3，表明变量分布的偏斜程度越严重；$S_k^{(1)}$的绝对值越接近于 0，表明变量分布的偏斜程度越轻微。

（二）利用四分位数求偏度系数

根据四分位数的特点可知，如果变量分布对称、无偏斜，那么第一个四分位数Q_L与第三个四分位数Q_U是关于中位数对称分布的，即$Q_U-m_e=m_e-Q_L$，因此我们可以通过$Q_U-m_e=m_e-Q_L$这个等式是否成立来判断变量分布是否对称，并且可以根据第一个、第三个四分位数与中位数距离的关系来求偏度系数并标记为$S_k^{(2)}$，计算公式为：

$$S_k^{(2)}=\frac{Q_L+Q_U-2m_e}{Q_U-Q_L} \tag{3-35}$$

偏度系数$S_k^{(2)}$的取值范围为（－1，1）。$S_k^{(2)}$的绝对值越接近于 1，表明变量分布的偏斜程度越严重；$S_k^{(2)}$的绝对值越接近于 0，表明变量分布的偏斜程度越轻微。

同理，我们也可以根据十分位数、百分位数来求偏度系数。

（三）利用动差法求偏度系数

计算偏度系数最重要的方法是动差法。动差法偏度系数是以变量数列的三阶中心动差（m_3）作为度量偏度的基本依据。动差又称为矩，原是物理学的概念，用以表示力与力臂对重心的关系。这个关系与加权算术平均数中变量值与权数对算术平均数的关系很相似，所以统计学上也用动差概念来说明变量分布的特征。

令常数 a 为变量分布的中心，则所有变量值与 a 值之差的 t 次方的算术平均数就称为变量 x 关于 a 的 t 阶动差，即：

$$t\text{阶动差}=\frac{\sum_{i=1}^{n}(x_i-a)^t}{n}$$ （根据未分组数据，可简记为 t 阶动差=$\frac{\sum(x_i-a)^t}{n}$）

（3-36）

或

$$t\text{阶动差}=\frac{\sum_{i=1}^{k}(x_i-a)^t f_i}{\sum_{i=1}^{k} f_i}$$ （根据变量数列，可简记为 t 阶动差=$\frac{\sum(x_i-a)^t f_i}{\sum f_i}$）

（3-37）

当 $a=0$ 时，t 阶动差称为 t 阶原点动差，若以 M_i 表示，则：

一阶原点动差为：$M_1=\frac{\sum x_i}{n}$ 或 $M_1=\frac{\sum x_i f_i}{\sum f_i}$，即算术平均数

二阶原点动差为：$M_2=\frac{\sum x_i^2}{n}$ 或 $M_2=\frac{\sum x_i^2 f_i}{\sum f_i}$，即平方的平均数

三阶原点动差为：$M_3=\frac{\sum x_i^3}{n}$ 或 $M_3=\frac{\sum x_i^3 f_i}{\sum f_i}$，即三次方的平均数

⋮

当 $a=\bar{x}$ 时，t 阶动差称为 t 阶中心动差，若以 m_i 表示，则：

一阶中心动差为：$m_1=\frac{\sum(x_i-\bar{x})}{n}$ 或 $m_1=\frac{\sum(x_i-\bar{x})f_i}{\sum f_i}$

二阶中心动差为：$m_2=\frac{\sum(x_i-\bar{x})^2}{n}$ 或 $m_2=\frac{\sum(x_i-\bar{x})^2 f_i}{\sum f_i}$

三阶中心动差为：$m_3=\frac{\sum(x_i-\bar{x})^3}{n}$ 或 $m_3=\frac{\sum(x_i-\bar{x})^3 f_i}{\sum f_i}$

⋮

很显然，一阶中心动差 $m_1=0$，偶数阶中心动差恒为正（其中 2 阶中心动差就是方差，即 $m_2=s^2$），而三阶及以上的奇数阶中心动差可正可负。由于变量分布的偏斜方向要通过偏度指标的正、负情况来反映，因此要用三阶及以上的奇数阶中心动差来衡量变量分布的偏斜方向。为了计算的方便，选择使用三阶中心动差 m_3 最

为合适。

当m_3=0 时，表示变量分布无偏；当m_3>0 时，表示变量分布是正偏；当m_3<0 时，表示变量分布为负偏。

由于m_3只是绝对数，因而不能直接比较。为了使不同变量分布的偏度比较具有相同的标准，就需要用相对数来衡量。我们把m_3与标准差的立方s^3对比，就得到了动差法的偏度系数，即：

$$S_k^{(3)}=\frac{m_3}{s^3} \qquad (3\text{-}38)$$

若$S_k^{(3)}$>0，表示变量分布正偏；若$S_k^{(3)}$<0，表示变量分布负偏；若$S_k^{(3)}$＝0，表示变量分布两边对称，无偏。$S_k^{(3)}$的绝对值越接近 0，表示变量分布的偏度越轻微；$S_k^{(3)}$的绝对值越大于 0，表示变量分布的偏度越严重；

【例 3-22】某企业职工月收入情况如表 3-15 所示，求职工月收入分布的动差法偏度系数。

表 3-15　某企业职工月收入情况表

职工月收入（元）	职工人数
900 以下	24
900～1000	48
1000～1100	60
1100～1200	105
1200～1300	27
1300～1400	21
1400～1500	12
1500 以上	3
合计	300

根据表 3-15 数据可得到动差法偏度系数计算表如表 3-16 所示。

表 3-16　某企业职工月收入动差法偏度系数计算表

职工月收入（元）	x_i	f_i	$x_i f_i$	$x_i-\bar{x}$	$(x_i-\bar{x})^2$	$(x_i-\bar{x})^2 f_i$	$(x_i-\bar{x})^3 f_i$

900 以下	850	24	20400	−263	69169	1660056	−436594728
900～1000	950	48	45600	−163	26569	1275312	−207875856
1000～1100	1050	60	63000	−63	3969	238140	−15002820
1100～1200	1150	105	120750	+37	1369	143745	5318565
1200～1300	1250	27	33750	+137	18769	506763	69426531
1300～1400	1350	21	28350	+237	56169	1179549	279553113
1400～1500	1450	12	17400	+337	113569	1362828	459273036
1500 以上	1550	3	4650	+437	190969	572907	250360359
合计		300	333900			6939300	404458200

根据表 3-16 数据可得：

$$\bar{x}=\frac{\sum x_i f_i}{\sum f_i}=\frac{333900}{300}=1113\text{（元）}$$

$$s=\sqrt{\frac{\sum (x_i-\bar{x})^2 f_i}{\sum f_i}}=\sqrt{\frac{6939300}{300}}=152.09\text{（元）}$$

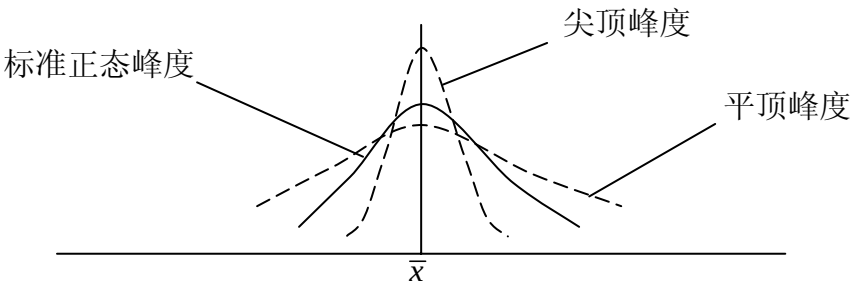
$$m_3=\frac{\sum (x_i-\bar{x})^3 f_i}{\sum f_i}=\frac{404458200}{300}=1348194\text{（元）}$$

$$S_k^{(3)}=\frac{m_3}{s^3}=\frac{1348194}{(152.09)^3}=0.38$$

结果表明，该企业职工月收入的分布为正偏分布，但偏度不大。

三、峰度系数

峰度的概念首先由统计学家皮尔逊于 1905 年提出，是对变量分布扁平性或尖陡性的测度，通常是指钟型分布的顶峰与标准正态分布相比偏扁平或偏尖陡的程度。它通常分为三种情况：标准正态峰度、尖顶峰度和平顶峰度，如图 3-6 所示。



3-6　变量分布不同峰度示意图

如果变量分布的频数比较集中于众数附近，分布曲线比较尖陡，使分布曲线的顶部较标准正态曲线更为突起，则变量分布的峰度属于尖顶峰度；如果变量分布各组的频数比较接近，分布曲线比较扁平，使分布曲线的顶部低于标准正态曲线，则变量分布的峰度属于平顶峰度。

峰度的测定是通过计算峰度系数来实现的，通常用 ***K*** 来表示。峰度系数的计

算主要采用动差法，是 4 阶中心动差 m_4 与标准差 4 次方 s^4 相比的结果，即：

$$K=\frac{m_4}{s^4} \qquad (3\text{-}39)$$

峰度系数的标准值为 3。当 $K=3$ 时，变量分布的峰度为标准正态峰度；当 $K<3$ 时，变量分布的峰度为平顶峰度；当 $K>3$ 时，变量分布的峰度为尖顶峰度。更进一步，当 K 值接近于 1.8 时，变量分布曲线就趋向于一条水平线，表示各组分配的频数接近于相同。当 K 值小于 1.8 时，则变量分布曲线为“U”型曲线，表示变量分布的频数分配是“中间少，两头多”。

【例 3-23】根据例 3-22 的表 3-15 数据计算职工月收入的峰度系数。

根据例 3-22 的有关计算结果可得：

$$m_4=\frac{\sum(x_i-\bar{x})^4 f_i}{\sum f_i}=1632660517\text{（元）}$$

$$s^4=535043161\text{（元）}$$

$$K=\frac{m_4}{s^4}=3.05$$

结果表明，该企业职工月收入分布的峰度为轻微的尖顶峰度。

本章小结

1. 变量分布特征的描述有以下三个方面：一是变量分布的集中趋势，反映变量分布中各变量值向中心值靠拢或聚集的程度；二是变量分布的离中趋势，反映变量分布中各变量值远离中心值的程度；三是变量分布的形状，反映变量分布的偏斜程度和尖陡程度。

2. 集中趋势亦称为趋中性，是指变量分布以某一数值为中心的倾向。作为中心的数值就称为中心值，它反映变量分布中心点的位置所在。变量分布的集中趋势要用平均指标来反映。平均指标是将变量的各变量值差异抽象化、以反映变量值一般水平或平均水平的指标，也就是反映变量分布中心值或代表值的指标。平均指标的具体表现称为平均数，平均数因计算方法不同可分为数值平均数和位置平均数两类。数值平均数主要包括算术平均数、调和平均数和几何平均数，位置平均数主要包括中位数和众数。在实际中，平均指标具有重要的作用。

3. 算术平均数也称为均值，是变量的所有取值的总和除以变量值个数的结果。根据数据的条件不同，有简单算术平均数与加权算术平均数之分。权数既表现为各组的频数，更表现为各组的频率。根据组距式数列计算的加权算术平均数是一个近似值。算术平均数具有两个重要的数学性质：各变量值与算术平均数的离差之和等于零，各变量值与算术平均数的离差平方之和为最小值。算术平均数易受极端值的影响。

4. 调和平均数从数学形式上看具有独立的形式，它是变量值的倒数的算术平均数的倒数，也称为倒数平均数。但在实际应用中，它更多地作算术平均数的变形而存在。调和平均数也有简单与加权之分，加权调和平均数的权数是各组的标志总量或各组标志总量占总体标志总量的比重。在以相对数或平均数计算平均数时，要能正确选择该使用加权算术平均数还是该使用加权调和平均数。

5.几何平均数是变量值的连乘积的相应次方根，是计算平均比率或平均速度的常用方法，例如用于计算水平法的平均发展速度、流水作业生产的产品平均合格率、复利法的平均利率等。它也有简单几何平均数和加权几何平均数两种。

6.从数学上看，算术平均数、调和平均数和几何平均数都是幂平均数的一种。

7.中位数是变量的所有变量值按定序尺度排序后，处于中间位置的变量值，是一种位置平均数。中位数既可用以测定定量变量的集中趋势，也可用以测定定序变量的集中趋势，但不适用于定类变量。分位数是将变量的数值按大小顺序排列并等分为若干部分后，处于等分点位置的数值。常用的分位数有四分位数、十分位数和百分位数。中位数就是一个特殊的分位数。

8.众数是变量数列中出现次数最多、频率最高的变量值，也是一种位置平均数。众数可用以测定任何种类变量的集中趋势。众数与中位数一样，都不受变量数列中极端值的影响。

9.利用算术平均数、众数、中位数三者之间的数量大小关系，可以判断变量分布是否对称以及偏斜的方向。在轻微偏斜时，可以根据已知的两个平均数去近似地估计第三个平均数。

10.所谓离中趋势，就是变量分布中各变量值背离中心值的倾向。变量分布的离中趋势要用离散指标来反映。离散指标就是反映变量值变动范围和差异程度的指标，即反映变量分布中各变量值远离中心值或代表值程度的指标，亦称为变异指标或标志变动度指标。离散指标具有重要的作用。常用的离散指标主要有：全距（亦称极差）、四分位差、异众比率、平均差、标准差、离散系数等，他们分别具有不同的特点与用途。方差和标准差具有若干重要的性质。

11.分布形状不同，表明变量分布的内在结构也不同。变量分布的形状要用形状指标来反映。形状指标就是反映变量分布具体形状，即左右是否对称、偏斜程度与陡峭程度如何的指标。形状指标有两个方面：一是反映变量分布偏斜程度的指标，称为偏度系数；二是反映变量分布陡峭程度的指标，称为峰度系数。计算偏度系数与峰度系数的主要方法是动差法。

练习与思考

一、判断题

1.对于定性变量，不能确定平均数。

2.根据组距式数列计算的平均数、标准差等，都是近似值。

3.任何平均数都受变量数列中的极端值的影响。

4.中位数把变量数列分成了两半，一半数值比它大，另一半数值比它小。

5.任何变量数列都存在众数。

6.如果$\bar{x} < m_e < m_0$，则变量分布为右偏。

7.若比较两个变量分布平均数代表性的高低，则方差或标准差大者平均数的代表性差。

8.只要变量分布具有相同的标准差，就会有相同的分布形状。

9.变量分布的集中趋势就是众数组的频数占总频数的比重，离中趋势则是非众数组的频数占总频数的比重。

10.在实际应用中，调和平均数与算术平均数的计算形式虽然不同，但计算结果及其意义是一样的。

二、单项选择题

1.由相对数计算平均数时，如果已知该相对数的子项数值，则应该采用(　　)。

A. 算术平均数　　B. 调和平均数
C. 几何平均数　　D. 位置平均数

2. 如果计算算术平均数的所有变量值都增加 100，则方差（　　）。

A. 增加 100　　B. 增加 10000　　C. 不变　　D. 不能确定如何变化

3. 如果计算加权算术平均数的各组频数都减少为原来的 4/5，则算术平均数（　　）。

A. 减少 4/5　　B. 减少为原来的 4/5　　C. 不变　　D. 不能确定如何变化

4. 某企业有 A、B 两个车间，去年 A 车间人均产量 3.6 万件，B 车间人均产量 3.5 万件。今年 A 车间生产人数增加 6%，B 车间生产人数增加 8%。如果两个车间的人均产量都保持不变，则该企业今年总的人均产量与去年相比（　　）

A. 上升　　B. 下降　　C. 不变　　D. 不能确定如何变化

5. 已知某变量分布属于钟型分布且 $m_0 = 900$，$m_e = 930$，则（　　）

A. $\bar{x} < 900$　　B. $900 < \bar{x} < 930$　　C. $\bar{x} > 930$　　D. $\bar{x} = 915$

6. 对某一变量数列计算数学意义上的数值平均数，得 $\bar{x} = 390$，则（　　）。

A. $H \geq 390, G \geq 390$　　B. $G \geq 390, H \leq 390$

C. $G \leq 390, H \geq 390$　　D. $G \leq 390, H \leq G$

7. 若两个变量数列的标准差相等且计量单位相同，但平均数不相等，则（　　）

A. 平均数大者代表性强　　B. 平均数小者代表性强
C. 两个平均数的代表性一样　　D. 无法判断哪个平均数的代表性强

8. 离散指标中受极端值影响最大的是（　　）

A. 平均差　　B. 标准差　　C. 全距　　D. 方差

9. 统计学中最重要的离散指标是（　　）

A. 平均差　　B. 全距　　C. 标准差　　D. 变异系数

10. 假如学生的考试成绩用优秀、良好、中等、及格和不及格来表示，那么全班考试成绩的水平高低应该用什么平均数来说明？（　　）

A. 可以用算术平均数　　B. 只能用众数
C. 可以用众数或中位数　　D. 只能用中位数

11. 根据动差的定义，方差属于（　　）

A. 一阶原点动差　　B. 二阶原点动差　　C. 一阶中心动差　　D. 二阶中心动差

12. 动差法峰度系数关于尖顶还是平顶的判断值是（　　）

A. 0　　B. 1　　C. 1.8　　D. 3

三、简答题

1. 什么是变量分布的集中趋势、离中趋势和分布形状？
2. 什么是平均指标？有什么作用？常用的平均数有哪些？
3. 如何理解加权平均数中权数的意义？试举例说明。
4. 在实际应用中，调和平均数与算术平均数有什么联系？
5. 从数学上看，算术平均数、几何平均数和调和平均数三者有什么关系？
6. 什么是中位数？有什么特点？试举例说明其应用。
7. 什么是众数？有什么特点？试举例说明其应用。
8. 算术平均数、中位数和众数三者的数量关系说明什么样的变量分布特征？
9. 什么是离散指标？有什么作用？常用的离散指标有哪些？

10.什么是方差和标准差？有哪些性质？

11.如何反映变量分布的形状？

四、计算题

1.某司机开车从A地到B地的时速是100公里，从B地返回A地的时速是120公里，问平均时速是多少？

2.菜场上某鱼摊大鲫鱼每条约重0.4公斤，售价为每公斤20元，小鲫鱼每条约重0.25公斤，售价为每公斤12元。某顾客向摊主提出大、小鲫鱼各买一条，一起称重，价格为每公斤16元。摊主应允，问这次买卖谁占了便宜？为什么？

3.某公司下属27家企业的资金利润率分组数据和各组年利润额数据如下表所示:

按资金利润率分组（%）	企业数	年利润额（万元）
8以下	2	300
8~12	6	1000
12~16	12	2600
16~20	5	1200
20以上	2	400
合 计	27	5500

请计算:

（1）平均每个企业的利润额;

（2）全公司的平均资金利润率（分别用绝对数权数和相对数权数）。

4.某年某企业3个车间的产品生产情况如下表所示:

车间	合格率（%）	合格品产量（辆）	年生产工时数（小时）
A	98	19600	6800
B	95	18620	7200
C	99	18434	8000
合计		56654	22000

问:

（1）若3个车间依次完成整辆产品某一工序的加工装配任务，全厂总的合格率、平均合格率和平均废品率分别是多少？

（2）若3个车间分别独自完成整辆产品的生产加工过程，则全厂总的合格率、平均合格率和平均废品率分别是多少？

（3）若3个车间生产的产品不同（使用价值不同），则全厂总的合格率、平均合格率和平均废品率又分别是多少？

5.甲班某次数学考试成绩如下表所示:

考试成绩（分）	学生人数
60以下	2
60~70	8
70~80	22
80~90	10
90以上	4
合 计	46

要求:

（1）计算算术平均数，四分位数和众数;

（2）计算全距，平均差，四分位差，异众比率，方差和标准差;

（3）计算偏度系数 $S_k^{(1)}$， $S_k^{(2)}$ 和 $S_k^{(3)}$;

（4）计算峰度系数;

（5）如果乙班的算术平均成绩为 80 分，标准差为 10 分，问哪个班级的平均成绩更有代表性?

6.某中学欲为初一 800 名新男生每人定制校服一套，小号、中号和大号三款分别适合身高 162cm 以下，162-168cm 和 168cm 以上的同学。根据以往数据知，初一男生的平均身高为 165cm，标准差为 3cm，问各款校服大概应分别准备多少套?

7.在定类变量中有一种叫两分类变量或是非变量，它只有两种结果，例如性别变量只有男或女两种结果。如果是非变量的两种结果分别用 1 和 0 来表示，那么该如何计算是非变量的平均数、方差、标准差和离散系数?请给出相关公式。

8.某班级 A、B、C 三门课程期末考试的平均成绩分别为 80 分、85 分和 88 分，标准差分别为 8 分、4 分和 7 分。甲、乙、丙三位同学该三门课程的考试成绩如下:

课 程 / 同 学	A	B	C
甲	77	91	89
乙	89	86	82
丙	69	93	95

问: 这三位同学的总分虽然都是 257 分，但实际上谁更具有竞争优势?

人物介绍

弗朗西斯.高尔顿（Francis Galton,1822-1911）：英国著名生物学家、统计学家，达尔文的近亲表弟。早年在剑桥大学学医，但医生职业对他并无吸引力。22 岁那年他获得一笔可观的遗产，决定弃医。1850 年至 1852 年，他与友人远赴非洲进行科学考察，1853 年被选为英国皇家地理学会会员，1856 年又被选为英国皇家学会会员。高尔顿研究涉猎范围包括地理、天文、气象、物理、机械、人类学、民族学、社会学、统计学、教育学、医学、生理学、心理学、遗传学、优生学、指纹学、照像术、登山术、音乐、美术、宗教等，是一位百科全书式的学者。主要著作有《气象测量》,《遗传的天才》,《自然的遗传》,《指纹》等 15 种，撰写各种学术论文 220 篇。高尔顿主张“无论何时，能算就算”，对统计学的最大贡献是相关性概念的提出和回归分析方法的建立。高尔顿的生物统计学思想经过他的学生皮尔逊、韦尔登的参与和发挥，在英国形成了一个颇有影响的生物统计学派。1901 年，高尔顿、皮尔逊、韦尔登创办《生物统计》杂志，成为生物统计学派的一面旗帜。1909 年，被英国王室授予勋爵称号。