3 第三章 数据分布特征的描述

通过调查获得并经过整理后展现的数据可以初步反映被研究对象的一些状态与特征，但认知程度还比较肤浅，反映的精确度不够。为此，我们需要使用各类代表性的数量特征值准确地描述这些数据。对单变量截面数据的特征描述主要有四个方面：集中趋势、离散程度、偏斜程度与峰度。

第一节 数据分布集中趋势的测度

集中趋势(central tendency)反映的是一组数据向某一中心值靠拢的倾向，在中心附近的数据数量较多，而远离中心的较少。对集中趋势进行描述就是寻找数据一般水平的中心值或代表值。根据取得这个中心值的方法不同，我们把测度集中趋势的指标分为两类：数值平均数和位置平均数。

一、数值平均数

数值平均数是同质总体内各个个体某一数量标志在一定时间、地点、条件下所达到的一般水平，是反映现象总体综合数量特征的重要指标，又称为平均指标。

研究总体中，各个个体的某个数量标志是各不相同的。例如，某个生产小组 10 名工人，如果按件取酬，则他们的工资分别是 1 000 元、1 480 元、1 540 元、1 600 元、1 650元、2 650元、2 740 元、2 800 元、2 900 元、3 500 元。要说明这 10 名工人工资的一般水平，显然不能用某一个工人的工资作代表，而应该计算他们的平均工资，用它作为代表值。

$$\text{平均工资}=\frac{1\,000+1\,480+\cdots+3\,500}{10}=2\,186\text{（元）}$$

2 186 元是在这 10 名工人工资的基础上计算出来的，其差异在计算过程中被抽象化了，结果得到的就是这 10 名工人工资的一般水平，即找到了一个代表值。

数值平均数有三种形式：算术平均数、调和平均数和几何平均数。

（一）算术平均数

算术平均数(arithmetic mean)是总体中各个体的某个数量标志的总和与个体总数的比值，一般用符号 $\bar{x}$ 表示。算术平均数是集中趋势中最主要的测度值，基本公式为

$$\text{算术平均数}=\frac{\text{某数量标志的总和}}{\text{对应的个体总数}}$$

由于所掌握的资料形式不同，算术平均数可以推导出两组公式。

▶ 1. 简单算术平均数

根据未经分组整理的原始数据计算算术平均数，设一组数据为 x_1，x_2，x_3，…，x_n，则其算术平均数为

$$\bar{x}=\frac{x_1+x_2+\cdots+x_n}{n}=\frac{\sum_{i=1}^{n}x_i}{n} \tag{3-1}$$

[例 3-1] 五名学生的身高分别为 1.65 米、1.69 米、1.70 米、1.71 米和 1.75 米，求他们的平均身高。

$$\text{解：}\bar{x}=\frac{\sum_{i=1}^{n}x_i}{n}=\frac{1.65+1.69+1.70+1.71+1.75}{5}=1.70(\text{米})$$

简单算术平均数之所以简单，就是因为各个变量值出现的次数相同，在例 3-1 中，每个变量值出现的次数都是 1，只要把各项变量值简单相加再用项数去除就求出平均数了。但是，当各变量值出现的次数不同时，情况就会发生变化。

▶ 2. 加权算术平均数

根据分组整理的数据计算平均数。设原始数据被分成 n 组，各组的变量值分别为 x_1，x_2，x_3，…，x_n，各组变量值出现的次数分别为 f_1，f_2，f_3，…，f_n，则这时的算术平均数为

$$\bar{x}=\frac{x_1f_1+x_2f_2+\cdots+x_nf_n}{f_1+f_2+\cdots+f_n}=\frac{\sum_{i=1}^{n}x_if_i}{\sum_{i=1}^{n}f_i} \tag{3-2}$$

计算加权算术平均数运用的变量数列资料有两种：单项变量数列和组距变量数列。单项变量数列直接对各组变量值进行加权平均计算即可；而组距变量数列需要先求出各组变量值的组中值，然后对组中值进行加权平均计算。

[例 3-2] 某车间 200 名工人加工零件的资料如表 3-1 所示，要求计算工人加工零件的平均数。

表 3-1 某车间工人加工零件平均数计算表

按零件数分组(个)	工人人数(人)f	人数比重	组中值 x	xf
40～50	20	0.10	45	900
50～60	40	0.20	55	2 200
60～70	80	0.40	65	5 200
70～80	50	0.25	75	3 750
80～90	10	0.05	85	850
合计	200	1.00	—	12 900

解：根据式(3-2)得

$$\bar{x}=\frac{\sum_{i=1}^{n}x_i f_i}{\sum_{i=1}^{n}f_i}=\frac{12\ 900}{200}=64.5(\text{个})$$

从以上计算过程可以看出次数 f 的作用：当变量值比较大的次数多时，平均数就接近于变量值大的一方；当变量值比较小的次数多时，平均数就接近于变量值小的一方。可见，次数对变量值在平均数中的影响起着某种权衡轻重的作用，因此被称为权数。

但是，如果各组的次数(权数)均相同时，即 $f_1=f_2=f_3=\cdots=f_n$ 时，则权数的权衡轻重作用也就消失了。这时，加权算术平均数会变成简单算术平均数，即

$$\bar{x}=\frac{\sum_{i=1}^{n}x_i f_i}{\sum_{i=1}^{n}f_i}=\frac{f\sum_{i=1}^{n}x_i}{nf}=\frac{\sum_{i=1}^{n}x_i}{n} \tag{3-3}$$

可见，简单算术平均数实质上是加权算术平均数在权数相等条件下的一个特例。

简单算术平均数数值的大小只与变量值的大小有关。加权算术平均数数值的大小不仅受各组变量值大小的影响，而且还受各组变量值出现的次数即权数大小的影响。

权数既可以用绝对数表示，也可以用相对数(比重)来表示。因此，加权算术平均数也可用式(3-4)的形式来表示。

$$\bar{x}=\sum_{i=1}^{n}x_i\frac{f_i}{\sum_{i=1}^{n}f_i} \tag{3-4}$$

[例 3-3] 仍以表 3-1 资料为例，当已知各组工人人数占全部工人人数的比重时，计算工人加工零件的平均数。

解：根据式(3-4)得

$$\begin{aligned}\bar{x}&=\sum_{i=1}^{n}x_i\frac{f_i}{\sum_{i=1}^{n}f_i}\\&=45\times0.1+55\times0.2+65\times0.4+75\times0.25+85\times0.05\\&=64.5(\text{个})\end{aligned}$$

针对统计调查获取资料的不同形式，计算评价指标的公式形式也不同，但可得到异曲同工的效果。用比重(频率)公式计算出来的平均指标与用绝对数次数作权数计算的结果是完全相同的。这是因为权数的两种形式，其计算公式在内容上是一样的。

▶ 3. 算术平均数的数学性质

算术平均数在统计学中有着重要的地位，它是进行统计分析和统计推断的基础，下面两个有关算术平均数的命题是算术平均数的两个重要的数学性质。

(1) 各变量值与其算术平均数离差之和等于零，即

$$\sum_{i=1}^{n}(x_i-\bar{x})=0 \tag{3-5}$$

式(3-5)证明如下：

$$\sum_{i=1}^{n}(x_i-\bar{x})=\sum_{i=1}^{n}x_i-\sum_{i=1}^{n}\bar{x}=\sum_{i=1}^{n}x_i-n\bar{x}=\sum_{i=1}^{n}x_i-n\cdot\frac{\sum_{i=1}^{n}x_i}{n}=\sum_{i=1}^{n}x_i-\sum_{i=1}^{n}x_i=0$$

(2) 各变量值与其算术平均数离差平方之和等于最小值，即

$$\sum_{i=1}^{n}(x_i-\bar{x})^2=\text{最小值}(\min) \tag{3-6}$$

式(3-6)证明如下：

设 x_0 为任意数，c 为常数($c\neq0$)，并令 $x_0=\bar{x}\pm c$，则

$$\begin{aligned}\sum_{i=1}^{n}(x_i-x_0)^2&=\sum_{i=1}^{n}[x_i-(\bar{x}\pm c)]^2=\sum_{i=1}^{n}[(x_i-\bar{x})\pm c]^2\\&=\sum_{i=1}^{n}(x_i-\bar{x})^2\pm2c\sum_{i=1}^{n}(x_i-\bar{x})+nc^2\\&=\sum_{i=1}^{n}(x_i-\bar{x})^2+nc^2\end{aligned}$$

因为 $nc^2>0$，所以

$$\sum_{i=1}^{n}(x_i-x_0)^2>\sum_{i=1}^{n}(x_i-\bar{x})^2$$

即 $\sum_{i=1}^{n}(x_i-\bar{x})^2$ 为最小值。

(二) 调和平均数

在统计分析中，有时会由于种种原因没有频数的资料，只有每组的变量值和相应的标志总量。这种情况下就不能直接运用算术平均方法来计算了，而需要以迂回的形式，即用每组的标志总量除以该组的变量值推算出各组的单位数，才能计算出平均数。这时，我们可以用调和平均的方法完成这个计算。

调和平均数(harmonic mean)是各变量值倒数的算术平均数的倒数。由于它是根据变量值倒数计算的，所以又称作倒数平均数，通常用 $\bar{x}_H$ 表示。根据掌握的资料不同，调和平均数可分为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

简单调和平均数是根据未经分组资料计算的调和平均数。我们先来看一个最简单的例子。

[例 3-4] 假如某种蔬菜在早、中、晚市场中，每市斤的单价分别为 1.5 元、1.4 元、1.2 元，若在早、中、晚市场中各买一市斤，其平均价格用简单算术平均数计算，结果是 1.37 元。但若早、中、晚市场各买一元钱，其平均价格是多少呢？

解：计算方法应先把总重量计算出来，然后再将总金额除以总重量，即

$$\text{平均价格}=\frac{\text{总金额}}{\text{总重量}}=\frac{1+1+1}{\frac{1}{1.5}+\frac{1}{1.4}+\frac{1}{1.2}}=\frac{3}{2.21}=1.35(\text{元/斤})$$

用公式表达为

$$\bar{x}_H=\frac{n}{\frac{1}{x_1}+\frac{1}{x_2}+\cdots+\frac{1}{x_n}}=\frac{n}{\sum_{i=1}^{n}\frac{1}{x_i}} \tag{3-7}$$

事实上，简单调和平均数是权数均相等条件下的加权调和平均数的特例。当权数不等时，就需要进行加权处理了。

▶ 2. 加权调和平均数

设 m 为加权调和平均数的权数，加权调和平均数公式为

$$\bar{x}_H = \frac{m_1 + m_2 + \cdots + m_m}{\frac{m_1}{x_1} + \frac{m_2}{x_2} + \cdots + \frac{m_n}{x_n}} = \frac{\sum_{i=1}^{n} m_i}{\sum_{i=1}^{n} \frac{m_i}{x_i}} \tag{3-8}$$

[例 3-5] 承上例，如果早、中、晚市场中购买蔬菜的金额不再是一元钱，而是如表3-2所示的情形，求购进的该种蔬菜的平均价格。

表 3-2 调和平均数计算表

时　　间	单价(元/斤)x	购买金额(元)m	购买量(斤)m/x
早市	1.5	4	2.67
中市	1.4	3	2.14
晚市	1.2	2	1.67
合计	—	9	6.48

解：根据式(3-8)计算蔬菜的平均价格为

$$\bar{x}_H = \frac{\sum_{i=1}^{n} m_i}{\sum_{i=1}^{n} \frac{m_i}{x_i}} = \frac{9}{6.48} = 1.40(\text{元})$$

▶ 3. 调和平均数是算术平均数的变形

调和平均数是算术平均数的变形，推导如下：

$$\bar{x}_n = \frac{\sum_{i=1}^{n} m_i}{\sum_{i=1}^{n} \frac{m_i}{x_i}} = \frac{\sum_{i=1}^{n} x_i f_i}{\sum_{i=1}^{n} \frac{x_i f_i}{x_i}} = \frac{\sum_{i=1}^{n} x_i f_i}{\sum_{i=1}^{n} f_i} = \bar{x}$$

调和平均数与算术平均数在本质上是一致的，不同之处在于统计调查获取的资料形式不同。在计算平均数时，应根据具体情况选择不同的公式。

(三) 几何平均数

几何平均数(geometric mean)是 n 个变量值连乘积的 n 次方根。几何平均数是计算平均比率和平均速度最适用的一种方法。通常用 $\bar{x}_G$ 表示。根据掌握的数据资料不同，几何平均数可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

简单几何平均数是根据未经分组资料计算的平均数，计算公式为

$$\bar{x}_G = \sqrt[n]{x_1 \cdot x_2 \cdot \cdots \cdot x_n} = \sqrt[n]{\prod_{i=1}^{n} x_i} \tag{3-9}$$

[例 3-6] 某产品生产需要经过六道工序，每道工序的合格率分别为 98%、91%、

93%、98%、98%、91%，求这六道工序的平均合格率。

解：因为成品的合格率等于各道工序产品合格率的连乘积，所以要用几何平均数来计算这六道工序的平均合格率，即

$$\bar{x}_G=\sqrt[6]{98\%\times91\%\times93\%\times98\%\times98\%\times91\%}=94.78\%$$

▶ 2. 加权几何平均数

当掌握的数据资料为分组资料，且各个变量值出现的次数不相同时，就要用加权方法计算几何平均数。加权几何平均数的公式为

$$\bar{x}_G=\sqrt[f_1+f_2+\cdots+f_n]{x_1^{f_1}\cdot x_2^{f_2}\cdot\cdots\cdot x_n^{f_n}}=\sqrt[\sum f]{\prod_{i=1}^{n}x_i^{f_i}} \tag{3-10}$$

[例 3-7] 某市从 2003—2016 年各年的工业增加值的增长率如表 3-3 所示，计算这 14 年的平均增长率。

表 3-3 某市一定时期的工业增加值的增长率

时 间	年 数	工业增加值的增长率(%)
2003—2006 年	4	10.2
2007—2011 年	5	8.7
2012—2016 年	5	9.6
合计	14	—

解：首先根据式(3-10)计算平均发展速度：

$$\bar{x}_G=\sqrt[f_1+f_2+\cdots+f_n]{x_1^{f_1}\cdot x_2^{f_2}\cdot\cdots\cdot x_n^{f_n}}$$

$$=\sqrt[4+5+5]{110.2\%^4\times108.7\%^5\times109.6\%^5}=109.45\%$$

再还原成平均增长率：

平均增长率＝平均发展速度－100%＝109.45%－100%＝9.45%

(四) 切尾均值

切尾均值(trimmed mean)为去掉大小两端的若干数值后计算中间数据的均值。这种方法在电视大奖赛、体育比赛及需要人们进行综合评价的比赛项目中已得到广泛应用，其计算公式为

$$\bar{x}_\alpha=\frac{x_{(n\alpha+1)}+x_{(n\alpha+2)}+\cdots+x_{(n+n\alpha)}}{n-2\times n\alpha}\qquad 0\leqslant\alpha\leqslant\frac{1}{2}$$

式中，n 为观察值的个数；α 为切尾系数。

[例 3-8] 某次歌手比赛共有 11 名评委，对某位选手的打分结果如下：

x_1	x_2	x_3	x_4	x_5	x_6	x_7	x_8	x_9	x_{10}	x_{11}
9.22	9.25	9.20	9.30	9.65	9.30	9.27	9.20	9.28	9.25	9.24

经整理得到顺序统计量值为

$x_{(1)}$	$x_{(2)}$	$x_{(3)}$	$x_{(4)}$	$x_{(5)}$	$x_{(6)}$	$x_{(7)}$	$x_{(8)}$	$x_{(9)}$	$x_{(10)}$	$x_{(11)}$
9.20	9.20	9.22	9.24	9.25	9.25	9.27	9.28	9.30	9.30	9.65

去掉一个最高分和一个最低分，取 $\alpha=1/11$，计算该选手的平均得分为

$$\bar{x}_{1/11}=\frac{x_{(11\times1/11+1)}+x_{(11\times1/11+2)}+\cdots+x_{(11-11\times1/11)}}{11-2\times11\times1/11}=\frac{x_{(2)}+x_{(3)}+\cdots+x_{(10)}}{11-2}$$

$$=\frac{9.2+9.22+\cdots+9.3}{9}=9.26$$

二、位置平均数

数值平均数是通过计算得出的反映变量集中趋势的指标，位置平均数则是通过定位找到的集中趋势指标。位置平均数包括中位数和众数两种。

（一）中位数

中位数是将一组数据按大小顺序排列后，处于中间位置的那个变量值，通常用 M_e 表示。其定义表明，中位数就是将某变量的全部数据均等地分为两半的那个变量值。其中，一半数值小于中位数，另一半数值大于中位数。中位数是一个位置代表值，因此它不受极端变量值的影响。

▶ 1. 未分组数据中位数的确定

对未分组数据资料，需先将各变量值按大小顺序排列，并按公式 $(n+1)/2$ 确定中位数的位置。当一个序列中的项数为奇数时，则处于序列中间位置的变量值就是中位数。例如，根据 7、6、8、2、3 这五个数据求中位数，先按大小顺序排成 2、3、6、7、8。在这个序列中，选取中间一个数值 6，小于 6 的数值有两个，大于 6 的数值也有两个，所以 6 就是这五个数值中的中位数。

当一个序列的项数是偶数时，则应取中间两个数的中点值作为中位数，即取中间两个变量值的平均数为中位数。例如一个按大小顺序排列的序列 2、5、7、8、11、12，其中位数的位置在 7 与 8 之间，中位数就是 7 与 8 的平均数，即 $M_e=(7+8)/2=7.5$。

▶ 2. 单项数列中位数的确定

根据单项数列资料确定中位数与根据未分组资料确定中位数的方法基本一致，首先计算各组的累计次数(或频数)，再按公式 $\left(\sum_{i=1}^{n}f_i+1\right)/2$ 确定中位数的位置，并对照累计次数确定中位数。

[例 3-9] 某班同学按年龄分组资料如表 3-4 所示，求中位数。

表 3-4 某班同学按年龄分组资料

年龄(岁)	学生人数(人)	向上累计	向下累计
17	5	5	50
18	8	13	45
19	26	39	37
20	9	48	11
21	2	50	2
合计	50	—	—

解：年龄中位数的位置为(50+1)/2=25.5，说明位于第 25 与第 26 位同学之间，根

据累计次数可确定中位数为第三组的变量值 19 岁。

▶ 3. 组距数列中位数的确定

根据组距数列资料确定中位数，应先按 $\left(\sum_{i=1}^{n} f_i\right)/2$ 的公式求出中位数所在组的位置，然后运用内插法按比例推算出中位数的近似值。

下限公式：

$$M_e = L + \frac{\frac{\sum_{i=1}^{n} f_i}{2} - S_{m-1}}{f_m} \tag{3-11}$$

上限公式：

$$M_e = U - \frac{\frac{\sum_{i=1}^{n} f_i}{2} - S_{m+1}}{f_m} \times i \tag{3-12}$$

式中，L 为中位数所在组的下限；U 为中位数所在组的上限；s_{m-1} 为向上累计至中位数所在组下一组（中位数组下限所在临组）的累计频数；s_{m+1} 为向下累计至中位数所在组上一组（中位数组上限所在临组）的累计频数；f_m 为中位数所在组的次数；i 为中位数所在组的组距。

上限公式和下限公式都是以中位数所在组内的次数均匀分布为前提的，在这种情况下才可以按比例推算中位数的近似值。

[例 3-10] 利用表 3-5 的资料，计算中位数。

表 3-5 中位数计算示例资料

按零件数分组(个)	职工人数(人)	向上累计	向下累计
40～50	20	20	200
50～60	30	50	180
60～70	90	140	150
70～80	50	190	60
80～90	10	200	10
合　　计	200	—	—

解：首先计算累计次数如表 3-5 所示，然后确定中位数所在组的位置为“60～70”组。将表 3-5 的资料代入中位数的上限公式和下限公式进行计算，所得结果一致。

按下限公式(3-11)计算：

$$M_e = L + \frac{\frac{\sum_{i=1}^{n} f_i}{2} - S_{m-1}}{f_m} \times i = 60 + \frac{\frac{200}{2} - 50}{90} \times 10 = 65.6(\text{个})$$

按上限公式(3-12)计算：

$$M_e = U - \frac{\frac{\sum_{i=1}^{n} f_i}{2} - S_{m+1}}{f_m} \times i = 70 - \frac{\frac{200}{2} - 60}{90} \times 10 = 65.6(\text{个})$$

从上面分析可知，中位数实际上就是位于累计次数达到 $\left(\sum_{i=1}^{n} f_i\right)/2$ 的这一组中的某个数值。该数值就是这一组下限加上按一定几何比例分割组距所得的一段组距，或这一组上限减去按一定几何比例分割组距所得的一段组距。有兴趣的同学可以根据图 3-1 对公式进行推导。

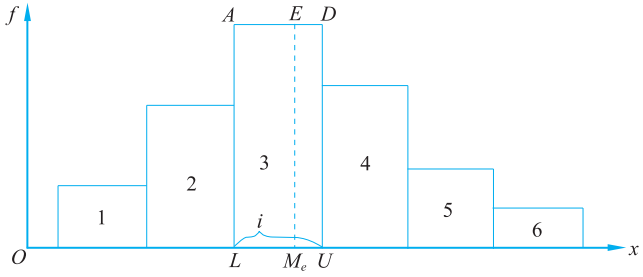


图 3-1 中位数计算公式推导示意图

提示：所谓中位数，就是有一半数据小于它，一半数据大于它，而直方图就是用面积表示次数的，所以以 EM_e 分界的两边面积应该相同。

▶ 4. 分位数

中位数是将统计分布从中间分成相等的两部分。同理，若将变量数列平均分成若干等份的等分点称为分位数。常用的分位数有四分位数、八分位数、十分位数和百分位数等。

三个数值可以将变量数列划分为项数相等的四部分，这三个数值就定义为四分位数(quartiles)，分别称为第一四分位数、第二四分位数和第三四分位数，记作 Q_1、Q_2 和 Q_3。对于不分组数据而言，三个四分位数的位置 Q_1 在 $(n+1)/4$ 处，Q_2 在 $(n+1)/2$ 处，Q_3 在 $[3(n+1)]/4$ 处，可见 Q_2 就是中位数。

同理，十分位数(decile)和百分位数(percentile)分别是将变量数列十等分和一百等分的数值。

(二) 众数

众数(mode)是一组数据中出现次数最多的那个变量值，通常用 M_o 表示。众数具有普遍性，在统计实践中，常利用众数来近似反映社会经济现象的一般水平。例如，说明某次考试学生成绩最集中的水平，或者说明城镇居民最普遍的生活水平等。

众数的确定要根据掌握的资料而定，未分组资料或单项数列资料众数的确定比较容易，不需要计算，可直接观察确定，即在一组数列或单项数列中，次数出现最多的那个变量值就是众数。如表 3-4 中，19 岁出现的人数最多，为 26 人，所以 19 岁就是众数。

根据组距数列确定众数比较复杂。首先要确定众数所在的组，若为等距数列，次数最多的那个组就是众数所在组；若为异距数列，需将其换算为次数密度(或标准组距次数)，换算后次数密度最多的一组即为众数所在组。然后按公式近似求出众数。

下限公式：

$$M_0 = L + \frac{f_m - f_{m-1}}{(f_m - f_{m-1}) + (f_m - f_{m+1})} \times i \tag{3-13}$$

上限公式：

$$M_0 = U - \frac{f_m - f_{m+1}}{(f_m - f_{m-1}) + (f_m - f_{m+1})} \times i \tag{3-14}$$

式中，L 为众数所在组的下限；U 为众数所在组的上限；i 为众数所在组的组距；f_m 为众数所在组的次数；f_{m-1} 为众数组下限所在临组的次数；f_{m+1} 为众数组上限所在临组的次数。

[例 3-11] 利用表 3-5 的资料，计算众数。

解：根据众数的概念确定众数所在组的位置为“60～70”组。将表 3-5 的资料代入众数的下限公式和上限公式，所得结果一致。

按下限公式(3-13)计算：

$$M_0 = L + \frac{f_m - f_{m-1}}{(f_m - f_{m-1}) + (f_m - f_{m+1})} \times i = 60 + \frac{90 - 30}{(90 - 30) + (90 - 50)} \times 10 = 66(\text{个})$$

按上限公式(3-14)计算：

$$M_0 = U - \frac{f_m - f_{m+1}}{(f_m - f_{m-1}) + (f_m - f_{m+1})} \times i = 70 - \frac{90 - 50}{(90 - 30) + (90 - 50)} \times 10 = 66(\text{个})$$

从上面的计算中可知，众数的数值要受到众数所在组相邻两组次数多少的影响。当众数组前一组次数大于众数所在组后一组次数时，众数接近众数组的下限；反之，当众数组前一组次数小于众数所在组后一组次数时，众数接近众数组的上限；而当众数所在组前后两组次数相等或当该数列次数呈对称分布时，众数所在组的组中值就是众数。有兴趣的同学可以根据图 3-2 对公式进行推导。

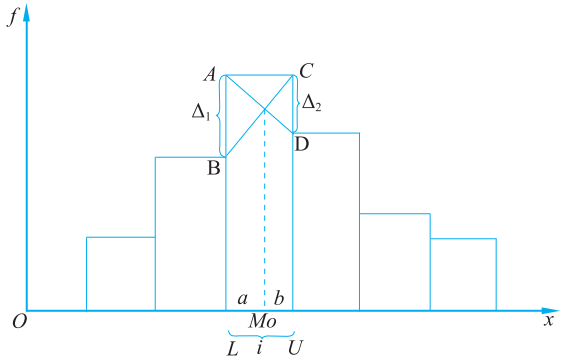


图 3-2　众数计算公式推导示意图

提示：用相似三角形原理证明，其中，$\Delta_1 = f_m - f_{m-1}$，$\Delta_2 = f_m - f_{m+1}$，$a = M_0 - L$，$b = U - M_0$。

（三）众数、中位数和算术平均数的比较

1. 众数、中位数和算术平均数的关系

大部分数据都属于单峰分布，其众数、中位数和算术平均数之间具有以下关系：

(1) 如果数据的分布是对称的，则 $M_0 = M_e = \bar{x}$，如图 3-3(a)所示。

(2) 如果数据是左偏分布，说明数据中偏小的数较多，这就必然拉动算术平均数向小

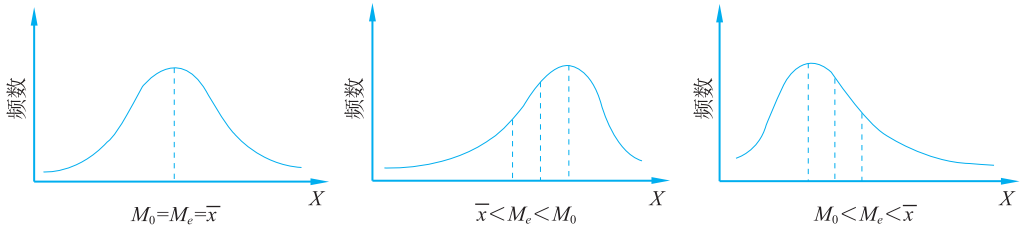


图 3-3 众数、中位数和算术平均数的关系

的一方靠，而众数和中位数由于是位置代表值，不受极端值的影响，因此三者之间的关系表现为 $M_0>M_e>\overline{x}$，又叫负偏，如图 3-3(b)所示。

(3) 如果数据是右偏分布，说明数据中偏大的数较多，必然拉动算术平均数向大的一方靠，则 $M_0<M_e<\overline{x}$，又叫正偏，如图 3-3(c)所示。

▶ 2. 众数、中位数和算术平均数的特点与应用场合

众数是一组数据分布的峰值，是位置代表值，其优点是易于理解，不受极端值的影响。当数据的分布具有明显的集中趋势时，尤其是对于偏态分布，众数的代表性比算术平均数要好。另外，众数不仅适用于定量测度水平的集中趋势分析，还适用于定类测度和定序测度水平的集中趋势评价。其特点是具有不唯一性，对于一组数据可能有一个众数，也可能有两个或多个众数，也可能没有众数。

中位数是一组数据中间位置上的代表值，其特点是不受极端值的影响。对于具有偏态分布的数据，中位数代表性要比算术平均数好。同时，中位数是定序测度变量集中趋势的最佳评价指标，但它不能用于定类测度水平的集中趋势度量。

算术平均数由全部数据的计算所得，具有优良的数学性质，是实际中应用最广泛的集中趋势测度值。其主要缺点是易受极端值的影响，对于偏态分布的数据，算术平均数的代表性较差。作为算术平均数变形的调和平均数和几何平均数是适用于特殊数据的代表值，调和平均数主要用于不能直接计算算术平均数的数据，几何平均数则主要用于计算比例数据的平均数，这两类平均数与算术平均数一样，易受极端值的影响。注意，数值平均数只适用于定量测度水平的数据，不能用于定类和定序数据。

在 SPSS 的 Frequencies 频数分析对话框中(见图 3-4)单击 Statistics 按钮，打开 Frequency: Statistics(统计量选择)对话框(见图 3-5)。图 3-5 中的 Central Tendency 选项列表中可选择集中趋势各项指标，包括算术平均数(Mean)、中位数(Median)、众数(Mode)和总和(Sum)。

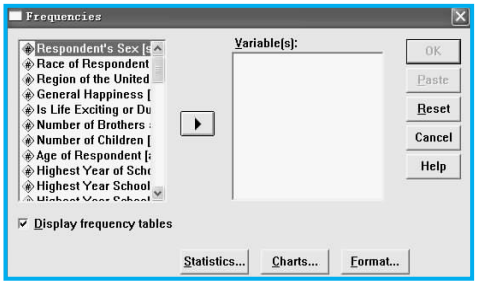


图 3-4 频数分析对话框

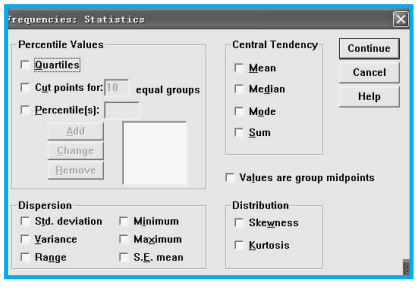


图 3-5 选择输出统计量对话框

第二节 离散程度的描述

集中趋势是一个说明同质总体各个体变量值的代表值，其代表性如何，取决于被平均的变量值之间的变异程度。在统计上，把反映现象总体中各个体之间差异程度的指标称为离散程度指标。反映离散程度的指标有绝对数和相对数两类。

一、离散程度的绝对指标

离散程度的绝对指标是用一个绝对数来反映总体中个体间的差异程度，主要包括极差、平均差、标准差等。

(一) 极差与分位差

▶ 1. 极差

极差(range)也叫变异全距，是一组数据的最大值与最小值之离差，即

$$R=\max(x_i)-\min(x_i) \tag{3-15}$$

式中，R 为极差；$\max(x_i)$和 $\min(x_i)$分别为一组数据的最大值和最小值。

对于组距分组数据，极差也可近似表示为

$$R\approx\text{最大组(变量值最大组)的上限值}-\text{最小组(变量值最小组)的下限值} \tag{3-16}$$

根据表 3-4，极差 $R=21-17=4$(岁)；根据表 3-1，极差 $R\approx90-40=50$(个)。

极差值越大说明总体中个体之间的差异越大。反之，极差值小则说明总体中个体间的差异也小。极差是描述数据离散程度的最简单测度值，它计算简单、易于理解，但只是说明两个极端变量值的差异范围，因此不能准确反映总体中个体间的差异程度，易受极端数值的影响。在企业的质量控制中，极差又称为“公差”，是对产品质量制定的一个容许变化的界限。

▶ 2. 分位差

分位差(divided difference)是计算剔除部分极端值后剩余数列的极差，是数列最大分位点与最小分位点之差。分位差是对极差指标的改进，常用的分位差有四分位差、八分位差、十分位差、十六分位差、三十二分位差以及百分位差等。

四分位差(interquartile range)是第三四分位数与第一四分位数之差，也称为内距或四分间距，用 Q_d 表示。四分位差的计算公式为

$$Q_d=Q_3-Q_1$$

四分位差反映了中间 50%数据的离散程度，其数值越小，说明中间的数据越集中；数值越大，说明中间的数据越分散。四分位差不受极端值影响，因此，在某种程度上弥补了极差的一个缺陷。

(二) 平均差

平均差(mean deviation)也称平均离差，是各变量值与其平均数离差绝对值的平均数，通常用 M_D 表示。由于各变量值与其平均数离差之和等于零，所以，在计算平均差时是取绝对值形式的。平均差的计算根据掌握数据资料的不同而采用两种不同的形式。

1. 简单式

对未经分组的数据资料，采用简单式，公式为

$$M_D = \frac{\sum_{i=1}^{n} | x_i - \bar{x} |}{n} \tag{3-17}$$

[例 3-12] 计算 5、11、7、8、9 的平均差。

解：先计算其算术平均数为 8，则代入式(3-17)得

$$M_D = \frac{| 5-8 | + | 11-8 | + | 7-8 | + | 8-8 | + | 9-8 |}{5} = 1.6$$

2. 加权式

根据分组整理的数据计算平均差，应采用加权式，公式为

$$M_D = \frac{\sum_{i=1}^{n} | x_i - \bar{x} | f_i}{\sum_{i=1}^{n} f_i} \tag{3-18}$$

[例 3-13] 承例 3-2 的资料，计算平均差。

解：根据例 3-2 的资料和结论 $\bar{x}=64.5$，将有关信息整理得表 3-6，代入式(3-18)中计算得：

$$M_D = \frac{\sum_{i=1}^{n} | x_i - \bar{x} | f_i}{\sum_{i=1}^{n} f_i} = \frac{1\,540}{200} = 7.7(\text{个})$$

表 3-6 平均差计算示例表

按零件数分组(个)	职工人数(人)f_i	组中值 x_i	$x_i-\bar{x}$	$\lvert x_i-\bar{x} \rvert f_i$
40～50	20	45	−19.5	390
50～60	40	55	−9.5	380
60～70	80	65	0.5	40
70～80	50	75	10.5	525
80～90	10	85	20.5	205
合　计	200	—	—	1 540

在可比的情况下，一般平均差的数值越大，则其平均数的代表性越小，说明该组变量值分布越分散；反之，平均差的数值越小，则其平均数的代表性越大，说明该组变量值分布越集中。

由于平均差是采用离差的绝对值形式进行数学计算，因此，在应用上有较大的局限性，因而不如标准差应用广泛。

(三) 标准差

标准差(standard deviation)又称均方差，它是各单位变量值与其平均数离差平方的均值的平方根，通常用 σ 表示。它是测量数据离散程度的最主要的方法。标准差具有量纲，

与变量值的计量单位相同。

标准差的本质是求各变量值与其平均数的距离和，即先求出各变量值与其平均数离差的平方，再求其平均数，最后对其开方。之所以称其为标准差，是因为在正态分布条件下，它和平均数有明确的数量关系，是真正度量离中趋势的标准。

根据掌握的数据资料不同，有简单式和加权式两种。

▶ 1. 简单式

对未经分组的数据资料，采用简单式，公式为

$$\sigma=\sqrt{\frac{\sum_{i=1}^{n}(x_i-\bar{x})^2}{n}} \tag{3-19}$$

[例 3-14] 计算 5、11、7、8、9 的标准差。

解：先计算其算术平均数为 8，则代入式(3-19)得

$$\sigma=\sqrt{\frac{(5-8)^2+(11-8)^2+(7-8)^2+(8-8)^2+(9-8)^2}{5}}=2$$

▶ 2. 加权式

根据分组整理的数据计算标准差，应采用加权式，公式为

$$\sigma=\sqrt{\frac{\sum_{i=1}^{n}(x_i-\bar{x})^2 f_i}{\sum_{i=1}^{n} f_i}} \tag{3-20}$$

[例 3-15] 承例 3-2 的资料，计算标准差。

解：根据例 3-2 的结论整理得表 3-7，代入式(3-20)可得 $\sigma=\sqrt{\frac{20\ 950}{200}}=10.23$(个)。

表 3-7 标准差计算示例表

按零件数分组(个)	职工人数(人)f_i	组中值 x_i	$x_i-\bar{x}$	$(x_i-\bar{x})^2$	$(x-\bar{x})^2 f_i$
40～50	20	45	−19.5	380.25	7 605
50～60	40	55	−9.5	90.25	3 610
60～70	80	65	0.5	0.25	20
70～80	50	75	10.5	110.25	5 512.5
80～90	10	85	20.5	420.25	4 202.5
合计	200	—	—	—	20 950

标准差是根据全部数据计算的，它反映了每个数据与其均值离差平方的平均数。因此，它能准确地反映出数据的离散程度。与平均差相比，标准差在数学处理上是通过平方消去离差的正负号，更便于数学上的处理。因此，标准差是实际中应用最广泛的离散程度测度值。

标准差有总体标准差与样本标准差之分，上述标准差公式是总体的标准差，样本标准差，需要在分母上减 1。一般用 S 表示样本标准差，其计算公式如下。

对简单式而言，

$$S=\sqrt{\frac{\sum_{i=1}^{n}(x_i-\bar{x})^2}{n-1}} \tag{3-21}$$

对加权式而言，

$$S=\sqrt{\frac{\sum_{i=1}^{n}(x_i-\bar{x})^2 f_i}{\sum_{i=1}^{n}f_i-1}} \tag{3-22}$$

方差(variance)是各变量值与其算术平均数离差平方和的平均数，即是标准差的平方，用 σ^2 表示总体方差，用 S^2 表示样本方差。

二、离散程度的相对指标

前面介绍的极差、平均差和标准差都是反映数据离散程度的绝对指标，其数据的大小一方面取决于原变量值的水平高低的影响，也就是与变量的平均数大小有关，变量值绝对水平高的，离散程度的测度值自然也就大，绝对水平低的，离散程度的测度指标自然也就小；另一方面，它们与原变量值的计量单位相同，采用不同计量单位计量的变量值，其离散程度的测度指标也就不同。

因此，对于平均数不等或计量单位不同的不同组别的变量值，就不能直接用离散程度的绝对指标来比较其离散程度。为了消除变量平均数不等和计量单位不同对离散程度的测度指标的影响，需要计算离散程度的相对指标，即离散系数，其一般公式为

$$\text{离散系数}=\frac{\text{离散程度的绝对指标}}{\text{对应的平均指标}}$$

离散程度通常是就标准差来计量的，因此，离散系数也称为标准差系数(coefficient of variation)，它是一组数据的标准差与其对应的平均数之比，是测度数据离散程度的相对指标，其计算公式为

$$V_\sigma=\frac{\sigma}{\bar{x}}\times 100\% \tag{3-23}$$

[例 3-16] 某地两个不同类型的企业全年平均月产量资料如表 3-8 所示，计算标准差系数。

表 3-8 企业平均月产量资料

企　　业	计 量 单 位	月平均产量 $\bar{x}$	标准差 σ
炼钢厂	吨	500	10
纺纱厂	锭	200	5

解：炼钢厂的标准差比纺纱厂大，却不能直接断定炼钢厂的平均月产量的代表性就比纺纱厂的小。这是因为：首先这两个厂的平均月产量相差悬殊，其次两个厂属于性质不同(计量单位不同)的两个企业，因此只能根据离散系数的大小来判断。根据式(3-23)得

$$\text{炼钢厂的离散系数 } V_\sigma=\frac{\sigma}{\bar{x}}\times 100\%=\frac{10}{500}=2\%$$

$$\text{纺纱厂的离散系数 } V_\sigma=\frac{\sigma}{\bar{x}}\times 100\%=\frac{5}{200}=2.5\%$$

两个企业的离散系数的计算结果表明，炼钢厂的平均月产量的代表性就比纺纱厂的大，生产比较稳定。其结果与用标准差判断的结果正好相反。

常用统计软件都有描述性统计功能，其中就包括各种离散程度指标。如图 3-5 所示的 Percentile Values 选项列表中给出了计算分位数选项：①四分位数(Quartils)，即 25%、50%、75%的位数；②任意分位数(Cut points for _ equal groups)；③百分位数[Percentile(s)]。Dispersion 选项列表中给出了变异指标选项：标准差(Std. Deviation)、方差(Variance)、极差(Range)、最小值(Minimum)、最大值(Maximum)、均值的标准误差(S. E. mean)。

三、数据的标准化

在计算了算术平均数和标准差之后，可以对一组数据中各个变量值进行标准化处理，以测度每个个体在总体中的相对位置，并可以用它来判断一组数据中是否存在异常值。标准化值是变量值与其平均数的离差与其标准差的比值，也称为 z 分数或标准分数。

设标准化数值为 z，则有

$$z=\frac{x_i-\bar{x}}{\sigma}\text{或}\ z=\frac{x_i-\bar{x}}{s} \tag{3-24}$$

[例 3-17] 如果几个学生的考试分数分别是 99、85、73、60、45、16，计算每位同学考试成绩的标准分数。

解：假定这次考试参加考试的所有学生考试成绩的算术平均数和标准差分别是：$\bar{x}=70.00$，$\sigma=15.00$。则根据式(3-24)计算这几位同学考试成绩的标准分数如下。

第一位同学考试成绩的标准分数为 $z=\frac{x_i-\bar{x}}{\sigma}=\frac{99-70}{15}=1.93$。

同理，计算出其他同学考试成绩的标准分数分别为 1.00、0.20、−0.67、−1.61、−3.60。

标准分数给出了一组数据中各数值的相对位置。例如，99 对应的标准分数为 1.93，显示出该生的考试成绩高于平均分 1.93 倍标准差。通常一组数据中高于或低于算术平均数三倍标准差的数值是很少的，即在算术平均数加减三个标准差的范围内几乎包含了全部数据，而在三倍标准差之外的数据，统计上称为离群点。例如，例3-17

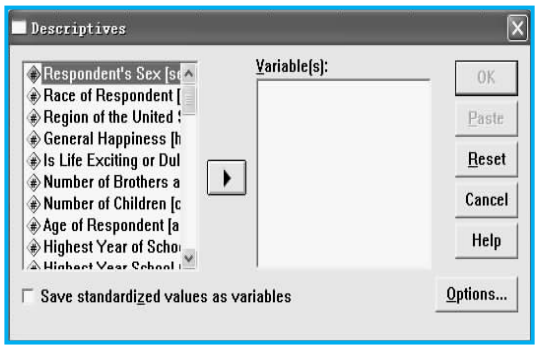


图 3-6　描述统计量分析对话框

中，考试成绩为 16 分的同学，其标准分数为－3.60，就是一个离群值，表明该生的考试成绩特别差。

标准化值在统计分析中的意义重大，统计软件的基本描述统计分析功能一般都能计算标准化值。SPSS 中标准化值的计算在描述统计量分析(Descriptives)对话框中，如图 3-6 所示。选中 Save standardized values as variables 复选项，即可对所选择的每一个变量进行标准化产生相应的 z 得分(标准化值)，作为新变量保存在数据窗中。

标准化后数据就没有量纲了，但不会改变其在原序列中的位置。在对多个具有不同量纲的变量进行比较分析时，常常需要对变量数值进行标准化处理。

四、是非标志标准差

是非标志是按照某一个品种标志，将总体划分为具有某一特征和不具有某一特征的两组。是非标志只有两种不同表现，可用 1 表示具有某一特征的标志，用 0 表示不具有某一特征的标志。总体的个体总数用 N 表示，具有某一特征标志的个体数用 N_1 表示，不具有某一特征标志的个体数用 N_0 表示，则 $N=N_1+N_0$，这两部分个体数占总体中的个体总数的比重可表示为

$$\pi=\frac{N_1}{N};\quad 1-\pi=\frac{N_0}{N}$$

π 是一个比率，它表示具有某种特征的个体的数量占总体中个体总数的比重，称为总体成数(或总体比率)。

是非标志的平均数为

$$\mu=\frac{\sum_{i=1}^{n}x_i f_i}{\sum_{i=1}^{n}f_i}=\frac{1\cdot\pi+0(1-\pi)}{\pi+(1-\pi)}=\pi \tag{3-25}$$

是非标志的标准差为

$$\begin{aligned}\sigma&=\sqrt{\frac{\sum_{i=1}^{n}(x_i-\bar{x})^2 f_i}{\sum_{i=1}^{n}f_i}}=\sqrt{\frac{(1-\pi)^2\cdot\pi+(0-\pi)^2(1-\pi)}{1}}\\&=\sqrt{\pi-2\pi^2+\pi^3+\pi^2-\pi^3}\\&=\sqrt{\pi(1-\pi)}\end{aligned} \tag{3-26}$$

从上述计算可以看出，是非标志的均值就是总体比率；是非标志的标准差就是具有某一特征标志的单位数在总体中的比重和不具有某一特征标志的单位数在总体中的比重两者的乘积的平方根(即两者的几何平均数)。

对于样本，样本容量用 n 表示，具有某一特征标志的个体数用 n_1 表示，不具有某一特征标志的个体数用 n_0 表示，则 $n=n_1+n_0$。若假定 p 为样本成数，则

$$\bar{x}=p \tag{3-27}$$

$$s=\sqrt{p(1-p)} \tag{3-28}$$

第三节 分布偏态与峰度的测度

集中趋势和离散程度是数据分布的两个重要特征，但要全面了解数据分布的特点，还需要掌握数据分布的形状是否对称、偏斜的程度以及扁平程度等。反映这些分布特征的测度指标是用来测度偏斜程度的偏态系数和用来测度平坦程度的峰度系数。

一、原点矩与中心矩

矩，又称为动差，来源于物理学中的“力矩”。物理学中的力矩用以测定转动趋势，说明某一点的作用力大小，它受作用力的大小和力臂的长度的影响。统计学中的“矩”是具有广泛意义的随机变量的数字特征。

（一）原点矩

以标志值 0 点为原点或支点，以各组标志值 x_i 为力臂的距离，以 $f_i/\sum_{i=1}^{n} f_i$ 为作用力的大小，则构成统计的一阶原点矩 u_1，即

$$\mu_1 = \frac{\sum_{i=1}^{n} x_i f_i}{\sum_{i=1}^{n} f_i} \tag{3-29}$$

如果将作用力臂分别采用各变量值的不同次方，如 x^2，x^3，…，x^k，则构成 k 阶原点矩，其一般式为

$$\mu_k = \frac{\sum_{i=1}^{n} x_i^k f_i}{\sum_{i=1}^{n} f_i} \tag{3-30}$$

（二）中心矩

若把原点移到算术平均数处，以 $x_i - \bar{x}$ 的各次方作为力臂的距离 $f_i/\sum_{i=1}^{n} f_i$ 为各作用力的大小，则构成统计的 k 阶中心矩 v_k，即

$$v_k = \frac{\sum_{i=1}^{n} (x_i - \bar{x})^k f_i}{\sum_{i=1}^{n} f_i} \tag{3-31}$$

在实际统计分析中，次数分布的一些统计特征值，如算术平均数和方差，可分别用一阶原点矩和二阶中心矩表示。在计算分布的特征状态偏斜度和峰度时，需要计算三阶、四阶原点矩和中心矩。

二、分布的偏态

偏态(skewness)是对分布偏斜方向和程度的测度。有些变量值出现的次数往往是非对

称型的，如收入分配、市场占有份额、资源配置等。变量分组后，总体中各个体在不同的分组变量值下分布并不均匀对称，而呈现出偏斜的分布状况，统计上将其称为偏态分布。

利用众数、中位数和平均数之间的关系就可以判断分布是对称、左偏还是右偏，但要测度偏斜的程度则需要计算偏态系数。统计分析中测定偏态系数的方法很多，一般采用矩的概念计算，其计算公式为三阶中心矩 v_3 与标准差的三次方之比，具体公式为

$$\alpha = \frac{v_3}{\sigma^3} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})^3 f_i}{\sum_{i=1}^{n} f_i \cdot \sigma^3} \tag{3-32}$$

式中，α 为偏态系数。

从式(3-32)可以看到，它是离差三次方的平均数再除以标准差的三次方。当分布对称时，离差三次方后正负离差可以相互抵消，因此 α 的分子等于 0，则 $\alpha=0$；当分布不对称时，正负离差不能抵消，就形成了正与负的偏态系数 α。当 α 为正值时，表示正偏离差值较大，可以判断为正偏或右偏；反之，α 为负值时，表示负偏离差值较大，可以判断为负偏或左偏。

偏态系数 α 的数值一般在 0 与 ± 3 之间，α 越接近 0，分布的偏斜度越小；α 越接近 ± 3，分布的偏斜度越大。

[例 3-18] 某管理局所属 30 个企业 2016 年 3 月的利润额统计资料如表 3-9 所示，要求计算该变量数列的偏斜状况。

表 3-9 偏斜系数计算示例表

利润额(万元)	企业数 f	组中值 x	$(x-\bar{x})^2 f$	$(x-\bar{x})^3 f$	$(x-\bar{x})^4 f$
10～30	2	**20**	**2 312**	**−78 608**	**2 672 672**
30～50	10	**40**	**1 960**	**−27 440**	**384 160**
50～70	13	**60**	**468**	**2 808**	**16 848**
70～90	5	**80**	**3 380**	**87 880**	**2 284 880**
合计	30	—	**8 120**	**−15 360**	**5 358 560**

解：利用表 3-9 中有关数据计算有关信息如表 3-9 中加粗部分数字。

根据式(3-20)计算标准差为

$$\sigma = \sqrt{\frac{\sum_{i=1}^{n}(x_i - \bar{x})^2 f_i}{\sum_{i=1}^{n} f_i}} = \sqrt{\frac{8\,120}{30}} = 16.45(\text{万元})$$

根据式(3-31)计算三阶中心距为

$$v_3 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})^3 f_i}{\sum_{i=1}^{n} f_i} = \frac{-15\,360}{30} = -512$$

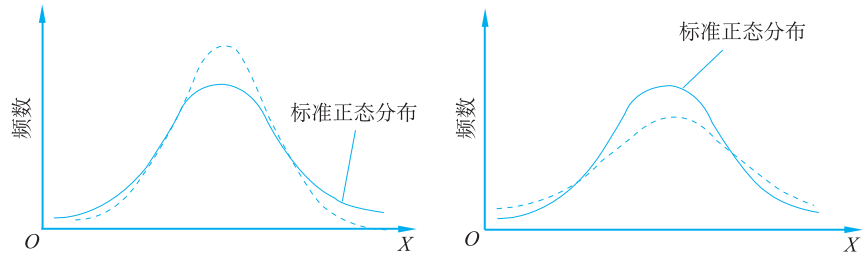
根据式(3-32)计算偏态系数为

$$\alpha = \frac{v_3}{\sigma^3} = \frac{-512}{16.45^3} = -0.12$$

计算结果表明该管理局所属企业利润额的分布状况呈轻微负偏分布。

三、分布的峰度

峰度(kurtosis)是分布集中趋势高峰的形状。在变量数列的分布特征中，常常以正态分布为标准，观察变量数列分布曲线顶峰的尖平程度，统计上称之为峰度。如果分布的形状比正态分布更高更瘦，则称为尖峰分布，如图 3-7(a)所示；如果分布的形状比正态分布更矮更胖，则称为平峰分布，如图 3-7(b)所示。



(a) 尖峰分布　　(b) 平峰分布

图 3-7　尖峰、平峰分布

测度峰度的方法，一般采用矩的概念计算，即运用四阶中心矩 v_4 与标准差的四次方之比，以此来判断各分布曲线峰度的尖平程度。公式为

$$\beta = \frac{v_4}{\sigma^4} - 3 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})^4 f_i}{\sum_{i=1}^{n} f_i \cdot \sigma^4} - 3 \tag{3-33}$$

式中，β 为峰度系数。

峰度系数是统计中描述次数分布状态的又一个重要特征值，测定邻近数值周围变量值分布的集中或分散程度。它以四阶中心矩为测量标准，除以 σ^4 是为了消除单位量纲的影响，而得到以无名数表示的相对数形式，以便在不同的分布曲线之间进行比较。正态分布的 $\frac{v_4}{\sigma^4}$ 值为 3，通常情况下将标准系数定义为 0 或 1，故峰度系数计算公式为 $\beta=\frac{v_4}{\sigma^4}-3$。则标准正态分布的峰度系数为 0，当 $\beta>0$ 时为尖峰分布，当 $\beta<0$ 时为平坦分布。

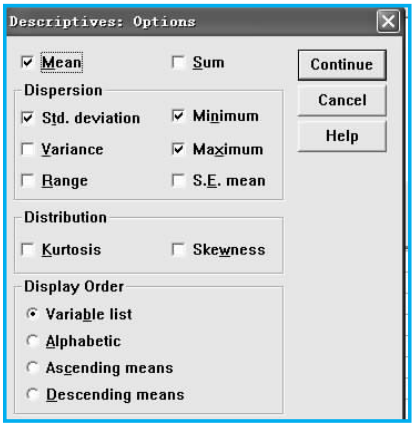


图 3-8　描述统计量分析对话框

[例 3-19] 承上例，计算该变量数列的峰度。

解：利用例 3-18 的计算结果和表 3-9 中有关数据计算峰度系数为

$$\beta = \frac{v_4}{\sigma^4} - 3 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})^4 f_i}{\sum_{i=1}^{n} f_i \cdot \sigma^4} - 3$$

$$= \frac{5\ 358\ 560}{30 \times 16.45^4} - 3 = 2.44 - 3 = -0.56$$

计算结果表明，上述企业间利润额的分布属于平坦分布(与标准正态分布比)，表明各变量值分布较为均匀。

数据的分布特征可以通过多个途径获得。例如，在 SPSS 中，频数分析的统计量(Frequencies Statistics)对话框(见图 3-5)和描述统计量分析(Descriptives Options)对话框(见图 3-8)都给出了分布特征选项(Distribution)，提供了偏态系数(Skewness)和峰度系数(Kurtosis)。

[例 3-20] 27 个工人分别看管机器的台数如下：

5 4 2 4 3 4 3 4 4 2 4 3 4 3 2 6 4 4 2 2 3 4 5 3 2 4 3

试分析计算这 27 人平均每人看管机器的台数。

要求：(1)指出这 27 人最多一人看管几台机器，最少看管几台机器？

(2)指出看管机器台数的众数、中位数和标准差指标。

(3)结合偏态系数和峰度系数说明看管机器台数的分布情况。

解：用 SPSS 的 Analyze→Descriptive Statistics→Frequencies 功能，选择机器台数作为分析变量后，单击 Statistics 按钮，打开统计量对话框，在 Dispersion 栏中选择 Std. Deviation、Maximum、Minimum；在 Central Tendency 中选择 Mean、Median、Mode；在 Distribution 中选择 Skewness 和 Kurtosis。

然后，在主对话框中单击 Charts 按钮，打开 Charts 对话框，选择 Histograms 项和 With normal curve 复选项，最后在主对话框中单击 OK 按钮。运行结果如图 3-9 和图3-10所示。

N	Valid	27
		0
Missing Mean		3.44
Median		4.00
Mode		4
Std. Deviation		1.050
Skewness		0.266
Std. Error of Skewness		0.448
Kurtosis		−0.119
Std. Error of Kurtosis		0.872
Minimum		2
Maximum		6

图 3-9 看管机器台数统计表

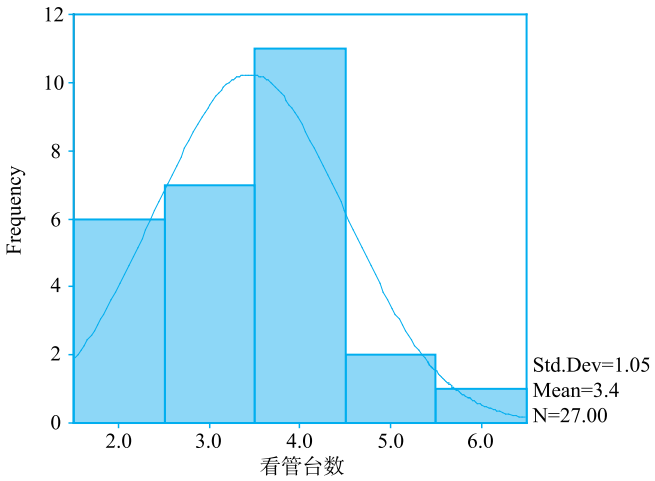


图 3-10 看管机器台数分布图

图 3-9 为不同看管台数的统计量。N 表示记录的个数其中分合法值的数量(valid)与缺失值的数量(missing)，各统计量的计算结果为 $\bar{x}=3.44$、$x_{max}=6$、$x_{min}=2$、$M_e=4$、$M_o=4$、$\sigma=1.05$、$\alpha=0.266$、$\beta=-0.119$。

从计算结果得知：$\alpha=0.266>0$(Skewness 为 0.266)，说明变量看管台数，数列分布为正偏(即右偏)有一个右尾；$\beta=-0.119<0$(Kurtosis 为 −0.119)，此值为负说明看管台数数列分布的峰度低于标准正态分布。该数列分布呈平坦的正偏(右偏)分布，这与图 3-10 显示的结果是一致的。

如果使用 Analyze→Descriptive Statistics→Descriptives 解决该问题的话，应在选好分析变量后单击 Options 按钮，在 Options 对话框中选择统计量 Mean，Dispersion 中的 Std. Deviation、Maximum 和 Minimum，Distribution 中的 Skewness 和 Kurtosis。最后在主对话框中单击 OK 按钮。运行结果如表 3-10 所示。结论与使用 Analyze→Descriptive Statistics→Frequencies 的分析结果一致。

表 3-10　使用 Descriptives 计算的看管台数基本统计量

	N	Minimu	Maxl	Mean	Std.	Skewness		Kurtosis	
	Statist	Statist	Statis	Statist	Statist	Statist	Std. Error	Statist	Std. Error
看管台数	27	2	6	3.44	1.050	0.266	0.448	−0.119	0.872
Valid N(listwise)	27								

本章要点

数据描述性指标主要分为两大类：一类测量数据的集中趋势；另一类测量数据的变异程度。在绝大多数情况下，特别是在正态分布情况下，掌握分布的集中趋势和离散程度就掌握了分布的所有特征。

反映分布的集中趋势的指标有三种：平均数、众数和中位数。在这三种指标中，平均数是最重要的，又细分为算术平均数、调和平均数和几何平均数。本章要求重点掌握算术平均数的计算形式和应用条件，即资料未分组的简单算术平均数的计算和分组后形成了变量数列(次数不等时)的加权算术平均数的计算。调和平均数是算术平均数的变形，应重点区分它与算术平均数的应用条件。几何平均数是计算平均发展速度等指标的重要方法，关于几何平均数应特别注意它的应用条件。从某种意义上讲，平均数是数据的平衡点。众数和中位数都是位置平均数，当数列中出现了极端值时，不宜采用算术平均数来表明现象的一般水平，此时应选择众数和中位数来表明现象的一般水平。

反映分布的离散程度的指标有多种，本章主要介绍了极差(变异全距)、分位差、标准差和变异系数。其中，变异全距(极差)和分位差在实际中较少应用，而标准差和方差是测量平均数代表性的重要指标，也是以后进行抽样估计的重要指标，应重点掌握。变异系数一般用于判断两个总体在平均数不等时的变异程度的比较。

均值(平均数)和方差一起反映了数据分布的位置和形状。本章最后介绍了反映分布形态的指标偏态系数和峰度系数，它们反映了数据的分布形态。

关键词

均值(mean)　　众数(mode)
几何平均数(geometric mean)　　中位数(median)
极差(range)　　方差(variance)
标准差(standard deviation)　　变异系数(cofficient of variation)
偏度(skewness)　　峰度(kurtosis)

思考题

1. 反映总体集中趋势的指标有哪几种？离散趋势指标有哪几种？并说明它们各自的特点和作用。

2. 在计算平均指标时，算术平均数、调和平均数和几何平均数分别适用于什么样的数据资料或应用条件？

3. 简单算术平均数和加权算术平均数的影响因素有哪些？

4. 数值平均数和位置平均数的区分依据是什么？两类平均数的应用有什么不同？

5. 说明算术平均数、众数和中位数三者的数量关系。

6. 标志变异指标有什么作用？

7. 变异系数与一般变异指标有什么区别？如何应用？

8. 偏度和峰度指标的作用及其与平均指标和变异指标的区别。

习　题

1. 某地区共有 50 万人，其中市区占 85%，郊区人口占 15%。为了了解该市居民的收入水平，在市区抽查了 1 500 户居民，每人年平均收入为 24 000 元；在郊区抽查了 1 000 户居民，每人年平均收入为 23 800 元。若这两个抽样数据均具有代表性，则该市居民年平均收入应为多少？

2. 某时期甲乙两农贸市场农产品交易资料如表 3-11 所示，计算并比较两市场农产品的平均价格。

表 3-11　甲乙两农贸市场农产品交易资料

品　　种	价格(元/千克)	甲市场成交额(元)	乙市场成交量(千克)
A	1.2	1.2	2
B	1.4	2.8	1
C	1.5	1.5	1
合计	—	5.5	4

3. 某车间有甲乙两个生产组，甲组平均每个工人日产量为 36 件，标准差为 9.6 件；乙组工人日产量资料如表 3-12 所示。

(1) 计算乙组平均每个工人的日产量和标准差。

(2) 比较甲乙两个生产组哪一组的日产量更具有代表性。

4. 某企业的工资资料如表 3-13 所示，计算该企业的职工平均工资。

表 3-12 乙组工人日产量

日产量(件)	工人数(人)
15	15
25	38
35	34
45	13

表 3-13 某企业的工资资料

工资水平(元)	职工比重(%)
3 000 以下	6
3 000～4 000	15
4 000～5 000	40
5 000～6 000	25
6 000 以上	14
合 计	100

5. 随机抽取 25 个网络用户，他们的年龄数据如下：

19 15 29 25 24
23 21 38 22 18
30 20 19 19 16
23 27 22 34 24
41 20 31 17 23

(1) 计算均值、众数和中位数。

(2) 计算四分位数。

(3) 计算标准差、偏态系数和峰度系数。

(4) 对网民的年龄分布特征进行综合分析。

6. 某公司下属三个企业的销售资料如表 3-14 所示，计算三个企业的平均利润率。

表 3-14 三个企业的销售资料

企 业	销售利润率(%)	销售额(万元)
甲	10	1 500
乙	12	2 000
丙	13	3 000

7. 某种产品的生产需要经过 10 道工序的流水作业，有 2 道工序的合格率是 90%，有 3 道工序的合格率是 92%，有 4 道工序的合格率是 94%，有 1 道工序的合格率是 98%，试计算该产品总的平均合格率。

8. 某酒店到三个农贸市场买草鱼，每千克的单价分别为 19 元、19.4 元、20 元，若各买 5 千克，则草鱼的平均价格是多少？若各买 100 元钱的，那么草鱼平均价格又是多少？

9. 某地区有一半家庭的月平均收入低于 600 元，一半高于 600 元，众数为 700 元，试估计算术平均数的近似值并说明其分布态势。

10. 有两个生产小组都有 5 个工人，某天，甲组的日产量件数为 8、10、11、13、15，乙组的日产量件数为 10、12、14、15、16。计算各组的算术平均数、全距、平均差、标准差和标准差系数，并说明哪组的平均数更具有代表性。

11. 某企业 6 月奖金如表 3-15 所示，计算算术平均数、众数、中位数并比较位置说明月奖金的分布特征。

12. 某地区农民的年收入资料如表 3-16 所示，计算农民年收入的众数和中位数。

表 3-15 某企业 6 月奖金

月奖金(元)	职工人数(人)
200～400	6
400～600	10
600～800	12
800～1 000	35
1 000～1 100	15
1 100～1 200	8
合　　计	86

表 3-16 某地区农民的年收入资料

按人均年收入分组(元)	户数(户)
20 000 以下	30
20 000～30 000	150
30 000～30 000	400
40 000～50 000	950
50 000～60 000	500
60 000～70 000	200
70 000～80 000	100
80 000 以上	50
合计	2 380

13. 对某企业甲乙两名工人当日产品中各抽取 10 件产品进行质量检查，产品的标准尺寸为 10mm，质量检查资料如表 3-17 所示，比较两名工人谁生产的零件尺寸更符合要求，谁的技术水平比较稳定。

14. 某笔投资收益的年利率资料如表 3-18 所示。

(1) 按复利计算，该笔投资收益的平均年利率为多少?

(2) 按单利计算，该笔投资收益的平均年利率为多少?

表 3-17 某企业产品质量检查资料

零件尺寸(mm)	零件数(件)	
	甲工人	乙工人
9.6 以下	1	1
9.6～9.8	2	2
9.8～10.0	3	2
10.0～10.2	3	3
10.2～10.4	1	2
合计	10	10

表 3-18 某笔投资收益的年利率资料

年利率(%)	年　　数
2	1
4	3
5	6
7	4
8	2

15. 2007—2016 年我国某城市职工平均工资和居民消费价格增长指数如表 3-19 所示，根据数据比较职工工资增长指数与平均居民消费价格指数的平均增长速度。

表 3-19 2007—2016 年我国职工工资和居民消费价格增长指数

年份	2007	2008	2009	2010	2011	2012	2013	2014	2015	2016
职工工资增长指数(%)	118.5	124.8	135.4	121.7	112.1	103.6	100.2	106.2	107.9	111.0
居民消费价格指数(%)	106.4	114.7	124.1	117.1	108.3	102.8	99.2	98.6	100.4	100.7

16. 某楼盘一年的出租情况如表 3-20 所示，计算平均租金和平均出租率。

表 3-20 某楼盘一年的出租情况

社 区	外租套数 x	出租率(%)f	租金(元)y
A	516	95	400
B	481	97	450
C	364	92	600
D	427	89	520

17. 气象局为研究我国的气温变化，对我国北方两个城市 1—2 月的气温做了记录，如表 3-21 所示。

表 3-21 我国北方两城市 1—2 月气温记录

气温(℃)	城市 A 的天数	城市 B 的天数
−30～−25	6	1
−25～−20	12	4
−20～−15	20	9
−15～−10	10	15
−10～−5	4	16
−5～0	3	7
0～5	3	4
5～10	1	3
合计	59	59

(1) 计算两城市的气温的均值。

(2) 计算两城市气温的标准差。

(3) 比较两城市气温离散程度的大小，并说明哪个城市的气候条件更适合居住。