

第三章 统计数据的描述性分析

第一节 相对程度的描述

第二节 集中趋势的描述

第三节 离中趋势的描述

第四节 分布形态的描述

第五节 运用SPSS进行统计量描述

2011年3月29日，时任国家统计局局长马建堂在《人民日报》发表文章，提醒人们“**全面认识我国在世界经济中的地位**”，其中有这样一组数字：2009年，我国三次产业就业人口在总就业人口中所占的比重分别为**38.1%、27.8%和34.1%**。这表明，我国从事第一产业即农业的人口比重过高，从事第二产业、第三产业的人口比重依然很低，其中从事第二产业的比重，仅相当于美国工业化初期即1870-1910年的水平，而第三产业更是比发达国家的比重低了一半。



- 不论是反映现象之间相对数量关系的相对指标，还是反映现象总体平均数量特征的平均指标，并不像直观所看到的两组数据的简单对比结果，数据的背后还有着复杂的社会背景以及指标运用过程中的一些局限；
- 本章将从**相对程度、集中趋势、离中趋势**开始，具体介绍统计数据的描述性分析方法。



第一节 相对程度的描述

- 一、相对指标的作用及表现形式
- 二、相对指标的种类
- 三、应用相对指标应注意的问题



一、相对指标的作用及表现形式

相对指标

又称作**相对数**，是通过两个具有一定联系的统计指标的对比，反映现象内部或现象之间的关联关系，即用数量对比的形式来表现总体的数量特征

相对指标的作用：

- 通过相互联系的现象数值之间的对比，反映事物间的数量联系程度（反映现象间的结构、比例、速度、程度和密度）
- 把现象的绝对差异抽象化，使一些不能直接对比的现象有了共同比较的基础



一、相对指标的作用及表现形式

相对指标的表现形式

无名数

构成相对指标分子和分母的两个指标的计量单位相同，表现为倍数、系数、成数、百分数（%）、千分数（‰）等

有名数

构成相对指标分子和分母的两个指标计量单位不相同，需要同时保留分子和分母的计量单位，也叫复名数，通常用来表示事物之间依存性的比例关系以及事物的密度和普遍程度等，比如人口密度用“人/平方公里”表示等



二、相对指标的种类

相对指标的种类

- 结构相对指标
- 比例相对指标
- 比较相对指标
- 强度相对指标
- 动态相对指标
- 计划完成程度相对指标



二、相对指标的种类

结构相对指标

在统计分组基础上，以总体总量作为比较标准，所求出的各组总量占总体总量的比重，用于反映总体中的部分数值与总体数值的结构比例关系

$$\text{结构相对指标} = \frac{\text{总体部分数值}}{\text{总体全部数值}} \times 100\%$$

- 一般用百分数或成数表示，其分子和分母可以是总体单位数，也可以是总体的标志数值；
- 由于是总体的部分数值与全部数值对比，结构相对数各部分所占比重之和必须等于100%或1。



二、相对指标的种类

例：下表是深圳股票市场不同板块上市公司的总数、总市值及其所占比重。其中上市公司比重是总体单位数的结构，总市值比重就是总体的标志数值的结构。

表 3-1 深圳股票市场总貌（2013 年 8 月 21 日）

深圳市场分类	上市公司（家）	上市公司比重（%）	总市值（亿元）	总市值比重（%）
主板市场	480	31.25	34409.39	42.01
中小板市场	701	45.64	34343.40	41.93
创业板	355	23.11	13152.53	16.06
合 计	1536	100.00	81905.32	100.00



二、相对指标的种类

比例相对指标

指总体中某一部分数值与总体的另一部分数值之比，用于反映总体中各组成部分之间所存在的比例关系

$$\text{比例相对指标} = \frac{\text{总体中某一部分数值}}{\text{总体中另一部分数值}}$$

- ▶ 进行连比时，根据总体多个部分的数值计算，一般将最小的部分设定为比较基础，以其为1或100%。



二、相对指标的种类

比较相对指标

指同一时间(或时点)的同类现象在不同总体(空间)之间的对比,说明该现象在不同空间上发展的不均衡程度

$$\text{比较相对指标} = \frac{\text{某一范围的指标数值}}{\text{另一范围的该指标数值}}$$

- 表现为无名数,可用倍数或百分数表示
- 对比的指标可以是绝对数,也可以是相对数或平均数



二、相对指标的种类

动态相对指标

指同一事物在不同时间上的数量对比关系。一般把研究的时期称为报告期，作为比较基础的时期则称为基期，用于说明同类事物在不同时间上的发展方向和程度。

$$\text{动态相对指标} = \frac{\text{报告期指标数值}}{\text{基期指标数值}} \times 100\%$$



二、相对指标的种类

强度相对指标

指两个性质不同但有一定联系的总量指标之比，用于反映现象发展的强度、密度和普遍程度

$$\text{强度相对指标} = \frac{\text{某一总体的指标数值}}{\text{另一有联系而性质不同总体的指标数值}}$$

- 表现形式一般为有名数，由指标的分子和分母的原有单位组成，如，人口密度用“人/平方公里”表示；
- 当分子和分母的计量单位相同时，则表现为无名数，一般用%或‰来表示，如，资金利税率用%表示，人口出生率、人口死亡率等用‰来表示。



二、相对指标的种类

- 一些强度相对指标的分子和分母位置可以互换，反映了现象在两种相反方向上的发展变化情况
- **正指标** 其数值大小与现象的强度、密度和普遍程度呈同向变化的强度相对数
 - **逆指标** 其数值大小与现象的强度、密度和普遍程度呈反向变化的强度相对数

已知某地区某年末总人口数为200万人，医院床位总数为48000张，则：

$$\text{每千人口拥有的医院床位数} = \frac{\text{医院床位数(张)}}{\text{人口数(千人)}} = \frac{48000(\text{张})}{2000(\text{千人})} = 24(\text{张/千人}) \quad (\text{正指标})$$

$$\text{每张医院床位负担的人口数} = \frac{\text{人口数(人)}}{\text{医院床位数(张)}} = \frac{2 \times 10^6}{48000} = 41.67(\text{人/张}) \quad (\text{逆指标})$$



二、相对指标的种类

计划完成程度相对指标

指某现象在某一段时间内的实际完成数值与计划任务数对比的结果

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

- 要求分子、分母在指标含义、计算口径、计算方法、计量单位以及时间长度和空间范围等方面完全一致。
- 由于计划总数是衡量计划完成情况的标准，故分子、分母不能互换。



三、应用相对指标应注意的问题

应用相对指标应注意的问题

- 注意指标之间的可比性问题
- 相对指标要与总量指标结合运用
- 多种相对指标结合运用



分地区年末人口数（单位：万人）

地 区	2009	2010	2011	2012	2013
全 国	133450	134091	134735	135404	136072
辽 宁	4341	4375	4383	4389	4390
吉 林	2740	2747	2749	2750	2751
黑龙江	3826	3833	3834	3834	3835

世界人口数（单位：百万人）

地 区	2009	2010	2011	2012	2013
中 国	1334.5	1340.9	1347.4	1354.0	1360.7
印 度	1198.0	1214.5	1241.5	1258.4	1252.1
美 国	314.7	317.6	313.1	315.8	320.1

地区生产总值（单位：百亿元）

地 区	2009	2010	2011	2012	2013
辽 宁	15.2	18.5	22.2	24.8	27.1
吉 林	7.3	8.7	10.6	11.9	13.0
黑龙江	8.6	10.4	12.6	13.7	14.4

第二节 集中趋势的描述

一、概述

二、定类数据：众数

三、定序数据：中位数和分位数

四、定距和定比数据：均值



一、概述

集中趋势

也叫**趋中性**，指总体中各数据值的次数分布从两边向中间集中的趋势

- 集中趋势反映各数据值向其中心值靠拢或聚集的程度；
- 度量集中趋势也就是寻找数据一般水平的代表值或中心值。



二、定类数据：众数

众数

指一组数据中出现次数最多的变量值，用 M_0 表示，主要用于测度定类数据的集中趋势

1.根据定类数据或单项式分组数列确定众数：

- 某基金投资的上市公司股票中有金融类2家、钢铁类3家、家电类2家、科技类8家、汽车类5家，在这组数据中出现次数最多的是科技类上市公司8家，因此，科技类公司即为这组定类数据的众数，即 $M_0 = \text{科技类公司}$



二、定类数据：众数

2.由定距数据（或组距数列）确定众数

- 需要根据分组数列中各组变量值出现次数的多少来确定，出现次数最多的组作为众数所在组，其所在组的区间作为众数的区间范围，再利用下列公式计算众数：

下限公式：

$$M_0 = L + \frac{f_m - f_{m-1}}{(f_m - f_{m-1}) + (f_m - f_{m+1})} \cdot d$$

上限公式：

$$M_0 = U - \frac{f_m - f_{m+1}}{(f_m - f_{m-1}) + (f_m - f_{m+1})} \cdot d$$



二、定类数据：众数

【例3.1】根据某日上海证券交易所790家上市公司股票价格的分组数据资料，确定众数的区间范围以及众数的具体数值。

某日上海证券交易所上市公司股票平均价格计算表

月股票价格 (元)	组中值(元) X	公司数(家) f	各组金额 (元) Xf
5以下	2.5	35	87.5
5-10	7.5	457	3427.5
10-20	15	268	4020
20以上	25	30	750
合计	—	790	8285



二、定类数据：众数

解：因为股票平均价格在5-10元的公司数量最多，共有457家，因此这一组即为众数所在的组，众数的区间范围为5-10元。

利用下限公式进一步计算众数：

$$\begin{aligned} M_0 &= L + \frac{f_m - f_{m-1}}{(f_m - f_{m-1}) + (f_m - f_{m+1})} \cdot d \\ &= 5 + \frac{457 - 35}{(457 - 35) + (457 - 268)} \times (10 - 5) = 5 + 3.45 = 8.45 \end{aligned}$$

或：利用上限公式进一步计算众数：

$$\begin{aligned} M_0 &= U - \frac{f_m - f_{m+1}}{(f_m - f_{m-1}) + (f_m - f_{m+1})} \cdot d \\ &= 10 - \frac{457 - 268}{(457 - 35) + (457 - 268)} \times (10 - 5) = 10 - 1.55 = 8.45 \end{aligned}$$



二、定类数据：众数

- 众数是衡量定类数据或品质数列位置的重要量度；
- 众数不受数列中各单位标志值的影响，当数列存在异常标志值时，能够较准确地代表总体各单位的一般水平；
- 若总体各单位变量值并没有明显的集中趋势，则众数不存在
- 如果在数据中恰好有两个众数，称此数据是双众数的；如果在数据中有多于两个的众数，则称此数据是多众数的。在多众数的情况下，对于描述数据的位置不能起多大作用，所以没有什么统计意义。



三、定序数据：中位数和分位数

中位数 是一组数据按照从小到大排序后，处于中间位置上的变量值，用 M_e 表示。

- 中位数将全部数据等分成两部分，每部分包含50%的数据，一部分数据比中位数大，另一部分数据则比中位数小。
- 是一种位置代表值，主要用于测度定序数据的集中趋势，也适用于定距数据和定比数据，但不适用于定类数据。
- 不受极端数值的影响，在总体标志值差异很大的情况下，具有较强的代表性。



三、定序数据：中位数和分位数

1.根据未分组资料确定中位数：

先将标志值按大小顺序排列，然后根据公式 $(N + 1)/2$ 来确定中位数的位置，再据此找到相对应的标志值

- 如果项数 N 为奇数，则中位数就是位于中央的那一项数值
- 如果项数 N 为偶数，则定义中位数为中央两项数值的平均数



三、定序数据：中位数和分位数

【例3.2】已知上海证券交易所七家上市公司的流通股本总数分别为（单位：亿股）：0.3，1.2，1.8，2.7，2.9，3.2，23，则有：

如果是八家上市公司，其流通股本总数分别为(单位：亿股)：0.3，1.2，1.8，2.7，2.9，3.2，23，28，则有：



三、定序数据：中位数和分位数

【例3.2】已知上海证券交易所七家上市公司的流通股本总数分别为（单位：亿股）：0.3，1.2，1.8，2.7，2.9，3.2，23，则有：

中位数的位次为： $\frac{N+1}{2} = \frac{7+1}{2} = 4$ ，即第4家上市公司的流通股本总数就是中位数： $M_e = 2.7$ 亿股

如果是八家上市公司，其流通股本总数分别为(单位：亿股)：0.3，1.2，1.8，2.7，2.9，3.2，23，28，则有：

中位数的位次为： $\frac{N+1}{2} = \frac{8+1}{2} = 4.5$ ，中位数应为第4和第5家上市公司流通股本的算术平均数，即

$$M_e = \frac{2.7 + 2.9}{2} = 2.8(\text{亿股})$$



三、定序数据：中位数和分位数

2.由单项式分组数列确定中位数：

因资料已经经过整理，并编制成标志值按大小顺序排列的变量数列，因此可直接用公式 $(\sum f + 1)/2$ 确定中位数的位次，再用向上累计或向下累计次数的方法，将累计次数超过中位数位次的组确定为中位数组，该组的标志值即为中位数。



三、定序数据：中位数和分位数

【例3.3】已知某只基金累计买入上海汽车股票278万股，其成交价格的分布情况如下，试确定该股票购买价格的中位数。

每股价格(元)	成交量(万股)	向上累计数(万股)	向下累计数(万股)
6.8	26	26	278
7.2	76	102	252
7.6	83	185	176
8.1	56	241	93
8.8	37	278	37
合计	278	—	—



三、定序数据：中位数和分位数

解：中位数的位次为：

$$\frac{\sum f + 1}{2} = \frac{278 + 1}{2} = 139.5(\text{万股})$$

从向上累计次数来看，第一组和第二组的累计次数为102万股，小于139.5万股，到第三组的累计次数为185万股，大于139.5万股，因此中位数必然在第三组，该组的标志值7.6元即为中位数。利用向下累计数确定中位数的方法与此相似，结果也相同。



三、定序数据：中位数和分位数

3.由组距式分组数列确定中位数：

从数列的累计频数栏中，找出第 $\sum f/2$ 个单位所在的组，即“**中位数组**”，该组的上、下限规定了中位数的可能取值范围；假定在中位数组内的各单位是均匀分布的，即可计算中位数的近似值：

下限公式
$$M_e = L + \frac{\frac{\sum f}{2} - S_{m-1}}{f_m} \times d$$

上限公式
$$M_e = U - \frac{\frac{\sum f}{2} - S_{m+1}}{f_m} \times d$$



三、定序数据：中位数和分位数

【例3.4】某银行按定期存款账号进行每隔10户的系统抽样，得到如下资料，试计算定期储蓄存款的中位数。

存款金额(元)	户数(户)	向上累计户数	向下累计户数
1000以下	30	30	500
1000-3000	150	180	470
3000-5000	250	430	320
5000-7000	50	480	70
7000-10000	10	490	20
10000以上	10	500	10
合计	500	—	—



三、定序数据：中位数和分位数

解：由于 $\frac{\sum f}{2} = \frac{500}{2} = 250$ ，从向上累计次数来看，第一组和第二组的累计次数均小于250户，到第三组的累计次数为430户，大于250户，由此可断定中位数落在第三组内。

根据下限公式计算中位数：

$$M_e = L + \frac{\frac{\sum f}{2} - S_{m-1}}{f_m} \times d = 3000 + \frac{\frac{500}{2} - 180}{250} \times 2000 = 3560(\text{元})$$

根据上限公式计算中位数：

$$M_e = U - \frac{\frac{\sum f}{2} - S_{m+1}}{f_m} \times d = 5000 - \frac{\frac{500}{2} - 70}{250} \times 2000 = 3560(\text{元})$$



三、定序数据：中位数和分位数

四分位数

把数列中各单位的变量值按照由小到大排序后，然后通过三个点将全部数据等分为四部分，其中每部分包含25%的数据，处于三个分割点位置上的数值就是四分位数（下四分位数、中位数、上四分位数）

根据未分组资料确定四分位数

- 首先对数据进行排序，然后再确定四分位数所在的位置；当四分位数的位置不在某一个数值上时，可以通过按比例分摊四分位数位置两侧数值的差值来估算。



三、定序数据：中位数和分位数

【例3.5】根据某只基金所购买的上海汽车股票价格分布表资料，试计算该基金所购买的上海汽车股票价格的上、下四分位数。

每股价格(元)	成交量(万股)	向上累计数(万股)	向下累计数(万股)
6.8	26	26	278
7.2	76	102	252
7.6	83	185	176
8.1	56	241	93
8.8	37	278	37
合计	278	—	—



三、定序数据：中位数和分位数

解：

下四分位数位置：
$$\frac{278 + 1}{4} = 69.75$$

上四分位数的位置：
$$\frac{3 \times (278 + 1)}{4} = 209.25$$

从表中的累计次数可以看到，下四分位数落在7.2元这一组中，上四分位数落在8.1元这一组中，即：

$Q_L=7.2$ 元，说明该股票的1/4或25%的购买价格在7.2元以下

$Q_U=8.1$ 元，说明该股票的3/4或75%的购买价格在8.1元以下



三、定序数据：中位数和分位数

百分位数

也是用于衡量数据位置的指标，提供有关各数据项如何在最小值与最大值之间分布的信息。
第 p 百分位的含义是，它使得至少有 $p\%$ 的数据项小于或等于这个值，且至少有 $(100-p)\%$ 的数据项大于或等于这个值

【例3.6】某班15名学生英语考试成绩按大小顺序排列后分别为：46，52，60，68，69，71，73，78，78，82，83，85，88，89，95，分别计算90%和60%的百分位数。



三、定序数据：中位数和分位数

解：

第90百分位数的位次： $i = (\frac{p}{100}) \times n = (\frac{90}{100}) \times 15 = 13.5$

由于*i*不是整数，将其向上取整，则第90百分位数的位置是大于13.5项的相邻整数，即第14项数据，其对应分数为89分，说明大约有90%的同学分数低于89分，有10%的同学分数高于89分。

第60百分位数的位次： $i = (\frac{p}{100}) \times n = (\frac{60}{100}) \times 15 = 9$

由于*i*是整数，则第60百分位数为第9项与第10项的平均值，即 $(78+82)/2=80$ ，说明大约有60%的同学分数低于80分，40%的同学分数高于80分。



四、定距和定比数据：均值

均值

也称为平均数，是集中趋势的主要测度指标，适用于定距数据和定比数据，不适用于定类数据和定序数据。根据所掌握数据的不同，均值有不同的计算形式和计算方法，其中最主要的表现形式是算术平均数。



四、定距和定比数据：均值

算术平均数

基本计算公式：

$$\frac{\text{总体中各单位某一数量标志之和}}{\text{总体单位总数}} = \frac{\sum m}{\sum f}$$

- 计算平均数时，分子与分母必须属于同一总体，且具有一一对应关系，即：有一个总体单位必须有一个标志值与之对应。
- 根据所掌握资料的表现形式不同，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形式。



四、定距和定比数据：均值

简单算术平均数

根据未分组资料计算，由各单位的标志值总和除以总体单位总数得到。

计算公式

$$\bar{X} = \frac{X_1 + X_2 + \cdots + X_N}{N} = \frac{\sum_{i=1}^N X_i}{N}$$

式中： $\bar{X}$ 为算术平均数； X_i 为第*i*个单位的标志值； N 为总体单位总数， Σ 表示求和符号。



四、定距和定比数据：均值

加权算术平均数

是根据分组数列资料计算的平均数

计算公式

$$\bar{X} = \frac{X_1 f_1 + X_2 f_2 + \cdots + X_m f_m}{f_1 + f_2 + \cdots + f_m} = \frac{\sum_{i=1}^m X_i f_i}{\sum_{i=1}^m f_i}$$

式中： f_i 为第*i*组的次数， m 为组数， X_i 为第*i*组的标志值或组中值。



四、定距和定比数据：均值

【例3.8】已知某投资者某日选中5只股票的价格和购买数量，试计算该投资者所购买股票的平均价格。

股票名称	价格 (元)	数量 (万股)	购买总金额 (万元)
某高速	3.5	16	56
某电子	4.1	12	49.2
某科技	5.6	18	100.8
某药业	9.8	2.5	24.5
某环保	15.6	1.8	28.08
合计	—	50.3	258.58



四、定距和定比数据：均值

解：

$$\bar{X} = \frac{\sum Xf}{\sum f} = \frac{3.5 \times 16 + 4.1 \times 12 + 5.6 \times 18 + 9.8 \times 2.5 + 15.6 \times 1.8}{16 + 12 + 18 + 2.5 + 1.8} = \frac{258.58}{50.3} = 5.14(\text{元})$$

说明：

- 上式是根据单项式分组数列计算的加权算术平均数，如果计算的依据是组距式分组数列，则需要用组中值代替各组标志值计算平均数，据此计算的平均数只是一个近似值。
- 用组中值作为各组标志值的代表值，其代表性如何，取决于各组内部的标志值分布是否呈均匀分布或对称分布。



四、定距和定比数据：均值

【例3.9】根据某日上海证券交易所上市公司股票平均价格的组距式分组数列，计算该交易所股票的平均价格。

月股票价格（元）	组中值(元) X	公司数(家) f	各组金额（元） Xf
5以下	2.5	35	87.5
5-10	7.5	457	3427.5
10-20	15	268	4020
20以上	25	30	750
合计	—	790	8285

解：
$$\bar{X} = \frac{\sum Xf}{\sum f} = \frac{2.5 \times 35 + 7.5 \times 457 + 15 \times 268 + 25 \times 30}{35 + 457 + 268 + 30} = \frac{8285}{790} = 10.49(\text{元})$$



四、定距和定比数据：均值

决定平均数的
变动范围

分析：

$$\bar{X} = \frac{\sum_{i=1}^m X_i f_i}{\sum_{i=1}^m f_i}$$

起到权衡轻
重的作用

权数

指各组标志值出现的次数，是变量值的承担者，反映了各组标志值对平均数的影响程度

绝对权数 表现为次数、频数、单位数，即 f_i

相对权数 表现为频率、比重，即 $f/\sum f$

$$\bar{X} = \frac{X_1 f_1 + X_2 f_2 + \cdots + X_m f_m}{f_1 + f_2 + \cdots + f_m} = X_1 \frac{f_1}{\sum f} + X_2 \frac{f_2}{\sum f} + \cdots + X_m \frac{f_m}{\sum f} = \sum X \cdot \frac{f}{\sum f}$$



四、定距和定比数据：均值

- 相对权数是计算加权算术平均数的实质权数，影响算术平均数大小的不是各组标志值出现次数的多少，而是各组单位数占总体单位数比重的高低。
- 若各组单位数相等，即 $f_1=f_2=\dots=f_m=A$ 时，则加权算术平均数与简单算术平均数的计算公式存在下面的关系式：

$$\bar{X} = \frac{\sum Xf}{\sum f} = \frac{\sum XA}{\sum A} = \frac{A\sum X}{mA} = \frac{\sum X}{m}$$

简单算术平均数是加权算术平均数的一个特例，是权数相等条件下的加权算术平均数。



四、定距和定比数据：均值

算术平均数的两个重要数学性质：

➤ 各单位标志值与算术平均数的离差之和等于零。即：

• 未分组资料 $\sum(X - \bar{X}) = 0$

• 分组资料 $\sum(X - \bar{X})f = 0$

➤ 各单位标志值与算术平均数的离差平方和最小。即：

• 未分组资料 $\sum(X - \bar{X})^2 = \min$

• 分组资料 $\sum(X - \bar{X})^2 f = \min$



四、定距和定比数据：均值

调和平均数

是总体各单位标志值倒数的算术平均数的倒数，也称倒数平均数

简单调和平均数计算公式

$$\overline{X}_H = \frac{1}{\frac{\frac{1}{X_1} + \frac{1}{X_2} + \cdots + \frac{1}{X_N}}{N}} = \frac{N}{\frac{1}{X_1} + \frac{1}{X_2} + \cdots + \frac{1}{X_N}} = \frac{N}{\sum \frac{1}{X}}$$

式中： $\overline{X}_H$ 表示调和平均数， X_i 表示第*i*组的标志值， N 表示总体单位总数。



四、定距和定比数据：均值

调和平均数

是总体各单位标志值倒数的算术平均数的倒数，也称倒数平均数

加权调和平均数计算公式

$$\overline{X_H} = \frac{1}{\frac{1}{X_1} f_1 + \frac{1}{X_2} f_2 + \cdots + \frac{1}{X_m} f_m} = \frac{f_1 + f_2 + \cdots + f_m}{\frac{1}{X_1} f_1 + \frac{1}{X_2} f_2 + \cdots + \frac{1}{X_m} f_m} = \frac{\sum f}{\sum \frac{1}{X} f}$$

式中： X_i 为第 i 组的标志值， f_i 为第 i 组的权数或次数， m 为组数。



四、定距和定比数据：均值

- 调和平均数一般作为算术平均数的变形来使用，即当已知总体标志总量和总体单位总数时，或者已知分组数列的各组标志值和标志总量时，可采用调和平均数计算平均指标
- 记各组的标志总量为 m 、各组的单位数为 f 、标志值为 X ，则有 $m = Xf$ ，当各组单位数未知时，原加权算术平均数公式变形为：

$$\bar{X} = \frac{\sum Xf}{\sum f} = \frac{\sum Xf}{\sum \frac{1}{X} \cdot Xf} = \frac{\sum m}{\sum \frac{1}{X} m}$$



四、定距和定比数据：均值

【例3.10】某投资者在不同时期内从上海证券交易所购买某上市公司的股票，购买价格分别为：23.56元, 15.70元和6.83元。如不计交易成本，试计算：

（1）如果在每一种价格水平上均购买了1000股，总平均价格是多少？

（2）如果在每一种价格水平上各购买了5万元，总平均价格是多少？



四、定距和定比数据：均值

解：

在每一种价格水平上均购买了1000股，总平均价格为：

$$\bar{X} = \frac{\sum Xf}{\sum f} = \frac{23.56 \times 1000 + 15.70 \times 1000 + 6.83 \times 1000}{1000 + 1000 + 1000} = 15.36(\text{元})$$

在每一种价格水平上各购买了5万元，总平均价格为：

$$\bar{X}_H = \frac{\sum m}{\sum \frac{1}{X} m} = \frac{50000 + 50000 + 50000}{\frac{50000}{23.56} + \frac{50000}{15.70} + \frac{50000}{6.83}} = 11.88(\text{元})$$



四、定距和定比数据：均值

几何平均数

N 项变量值连乘积的开 N 次方根，用于计算现象的平均比率或平均速度

应用的前提条件：

- 各个比率或速度的连乘积等于总比率或总速度
- 相乘的各个比率或速度不为零或负值

简单几何平均数（适用于未分组资料）

$$\bar{X}_G = \sqrt[N]{X_1 \cdot X_2 \cdots X_N} = \sqrt[N]{\prod X}$$

式中： N 为变量值的个数， X_i 为第 i 个变量值。



四、定距和定比数据：均值

【例3.11】某流水生产线有前后衔接的五道工序。某日各工序产品的合格率分别为95%、92%、90%、85%、80%，求整个流水生产线产品的平均合格率。

解：因该流水线最终合格品即为第五道工序的合格品，故该流水线总的合格品应为：

$$100A \times 0.95 \times 0.92 \times 0.90 \times 0.85 \times 0.80$$

则该流水线产品总的合格率为：

$$\begin{aligned} \frac{\text{总合格品}}{\text{总产品}} &= \frac{100A \times 0.95 \times 0.92 \times 0.90 \times 0.85 \times 0.80}{100A} \\ &= 0.95 \times 0.92 \times 0.90 \times 0.85 \times 0.80 \end{aligned}$$



四、定距和定比数据：均值



即：该流水线总的合格率等于各工序合格率的连乘积，符合几何平均数的适用条件，则

$$\begin{aligned}\overline{X}_G &= \sqrt[5]{0.95 \times 0.92 \times 0.90 \times 0.85 \times 0.80} \\ &= \sqrt[5]{0.5349} = 88.24\%\end{aligned}$$

思考

若上题中不是由五道连续作业的工序组成的流水生产线，而是五个独立作业的车间，且各车间的合格率同前，又假定各车间的产量相等均为100件，求该企业的平均合格率。



四、定距和定比数据：均值

分析：

因各车间彼此独立作业，所以有

第一车间的合格品为： $100A \times 0.95$ ；

第二车间的合格品为： $100A \times 0.92$ ；

.....

第五车间的合格品为： $100A \times 0.80$ 。

则该企业全部合格品应为各车间合格品的总和。

即：总合格品= $100A \times 0.95 + \dots + 100A \times 0.80$



四、定距和定比数据：均值

不再符合几何平均数的适用条件，需按照求解比值的平均数的方法计算。又因为

$$\text{合格率}(X) = \frac{\text{合格品}(m)}{\text{产品}(f)}$$

应采用加权算术平均数公式计算，即

$$\begin{aligned}\bar{X} &= \frac{\sum Xf}{\sum f} = \frac{0.95 \times 100A + \cdots + 0.80 \times 100A}{100A + \cdots + 100A} \\ &= \frac{442}{500} = 88.4\%\end{aligned}$$



四、定距和定比数据：均值

加权几何平均数（适用于分组资料）

$$\overline{X}_G = \sum_{i=1}^m f_i \sqrt[f_i]{X_1^{f_1} \cdot X_2^{f_2} \cdots X_m^{f_m}} = \sum_{i=1}^m f_i \sqrt[m]{\prod_{i=1}^m X_i^{f_i}}$$

式中： X_i 为第 i 组的标志值, f_i 为第 i 组的次数； m 为组数, $\prod$ 表示连乘符号。



四、定距和定比数据：均值

【例3.12】某金融机构以复利计息。近12年来的年利率有4年为3%，2年为5%，2年为8%，3年为10%，1年为15%。求平均年利率。

分析：按照复利计息，则该金融机构12年总本利率等于各年本利率的连乘积，符合几何平均数的适用条件，故计算平均年本利率应采用几何平均法。

解：

$$\begin{aligned}\bar{X}_G &= \sqrt[4+2+\cdots+1]{(1+0.03)^4(1+0.05)^2\cdots(1+0.15)} \\ &= \sqrt[12]{2.2154} = 106.85\%\end{aligned}$$

平均年利率为： $\bar{X}_G - 1 = 106.85\% - 1 = 6.85\%$



四、定距和定比数据：均值

思考

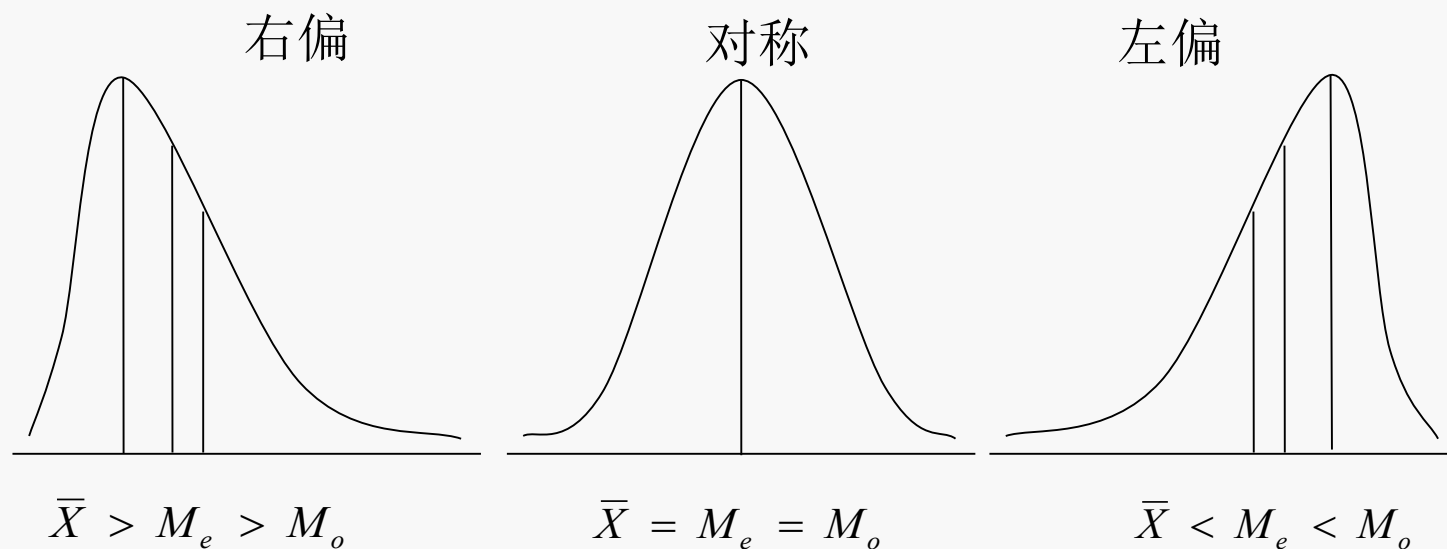
若上题中不是按复利而是按单利计息，且各年的利率与上相同，那么应该如何求平均年利率？这时，利息率或本利率不再符合几何平均数的适用条件，需要采用加权算术平均数公式计算平均年利息率。

设本金为V，则平均年利率的计算过程如下：

$$\begin{aligned}\bar{X} &= \frac{\sum Xf}{\sum f} = \frac{(V \times 0.03) \times 4 + \cdots + (V \times 0.15) \times 1}{V \times 4 + \cdots + V \times 1} \\ &= \frac{0.83V}{12V} = 6.92\%\end{aligned}$$



第二节 集中趋势的描述



- 众数、中位数和平均数是反映集中趋势的三个主要测度指标，各自具有不同的特点和应用场合。
- 从分布的角度来看，众数始终是一组数据分布的最高峰值，中位数是处于一组数据中间位置上的值，而平均数则是全部数据的算术平均。



第二节 集中趋势的描述

- **众数**的优点是不受极端值的影响，缺点是具有不唯一性。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。主要适合于作为分类数据的集中趋势的测度指标。
- **中位数**和**其他分位数**的特点也是不受数据极端值的影响，主要适合于作为顺序数据的集中趋势测度值。
- **平均数**针对数值型数据，并且利用了全部数据信息，是实践中应用最广泛的集中趋势的测度指标，其缺点主要是易受数据极端值的影响。



第二节 集中趋势的描述

- 当数据呈**对称分布**或接近对称分布时，众数、中位数和平均数将接近相等，这时应选择**平均数**作为集中趋势的测度指标。
- 对于偏态分布的数据，特别是当偏斜的程度较大时，可以考虑选择**众数或中位数**等位置平均数作为集中趋势的代表值，这时它们的代表性要比平均数好。



第三节 离中趋势的描述

- 一、定类数据：异众比率
- 二、定序数据：四分位差
- 三、定距定比数据：标准差和方差
- 四、离散系数



第三节 离中趋势的描述

- 数据的离散程度是数据分布的另一个重要特征，反映的是各变量值远离其中心值的程度，也称为离中趋势。
- 数据的离散程度越大，集中趋势的测度指标对该组数据的代表性就越差，离散程度越小，其代表性就越好。
- 根据所依据数据类型的不同，描述数据离散程度采用的测度指标主要有：异众比率、四分位差、方差和标准差、极差、平均差以及测度相对离散程度的离散系数等。



一、定类数据：异众比率

异众比率

指非众数组的频数占总频数的比率，用于衡量众数对一组数据的代表程度

计算公式：

$$V_r = \frac{\sum f_i - f_m}{\sum f_i} = 1 - \frac{f_m}{\sum f_i}$$

式中： f_m 为众数组的频数； $\sum f_i$ 为变量值的总频数

- 异众比率越大，非众数组的频数占总频数的比重越大，表明众数的代表性越差；异众比率越小，表明众数的代表性越好。



一、定类数据：异众比率

【例3.13】已知某基金投资的上市公司股票中有金融类2家、钢铁类3家、家电类2家、科技类8家、汽车类5家，试计算异众比率。

解：由于科技类上市公司最多为8家，即众数 $M_0=8$ ，进一步计算该组定类数据的异众比率：

$$V_r = \frac{20 - 8}{20} = 1 - \frac{8}{20} = 60\%$$

表明基金所购买的20家上市公司股票中，非科技类上市公司占60%，异众比率相对较大，为此用科技类上市公司代表该基金投资品种的状况，其代表性并不高。



二、定序数据：四分位差

四分位差

指上四分位数与下四分位数之差，也称为内距或四分间距。

计算公式：

$$Q_d = Q_u - Q_l$$

- 反映了中间50%数据的离散程度，其数值越小，说明中间的数值越集中；数值越大，说明中间的数据越分散。
- 不受极端数值的影响，在一定程度上也可以说明中位数对一组数据的代表程度。



三、定距定比数据：标准差和方差

极差 也称全距，指总体中各单位标志值最大值与最小值之差，通常用来说明总体中标志值变动的范围。

计算公式：

$$R = X_{\max} - X_{\min}$$

最大变量值或最高组上限或开口组假定上限

最小变量值或最低组下限或开口组假定下限

- 极差越大，表明总体中标志值的变动范围越大，说明总体各单位标志值的差异越大。
- 极差越小，总体中标志值的变动范围越小，说明总体各单位标志值的差异越小。



三、定距定比数据：标准差和方差

极差的特点：

- 优点：计算方法简单、易懂。
- 缺点：易受极端数值的影响，不能全面反映所有标志值差异大小及分布状况，准确程度差。
- 在生产实践中，人们一般是根据极差与标准差的关系，应用它来绘制产品的质量控制图，并配合抽样方法来检验产品质量，从而实现对产品的质量控制。



三、定距定比数据：标准差和方差

平均差

是各单位标志值对其算术平均数离差绝对值的平均数，反映了各单位标志值对其算术平均数的平均离差。平均差越大，说明各单位标志值的差异程度越大；反之，则说明其变动程度越小

(1) 对于未分组资料

$$A \cdot D = \frac{\sum_{i=1}^N |X_i - \bar{X}|}{N}$$

(2) 对于分组资料

$$A \cdot D = \frac{\sum_{i=1}^m |X_i - \bar{X}| f_i}{\sum_{i=1}^m f_i}$$



三、定距定比数据：标准差和方差

【例3.18】计算某日上海证券交易所790家上市公司股票价格的平均差。

股票价格 (元)	组中值 (元) X	公司数 (家) f	$ X - \bar{X} $	$ X - \bar{X} f$
5以下	2.5	35	7.99	279.65
5-10	7.5	457	2.99	1366.43
10-20	15	268	4.51	1208.68
20以上	25	30	14.51	435.3
合计	—	790	—	3290.06



三、定距定比数据：标准差和方差

解：

$$\bar{X} = \frac{\sum Xf}{\sum f} = \frac{2.5 \times 35 + 7.5 \times 457 + 15 \times 268 + 25 \times 30}{35 + 457 + 268 + 30} = \frac{8285}{790} = 10.49(\text{元})$$

$$A \cdot D = \frac{\sum_{i=1}^m |X_i - \bar{X}| f_i}{\sum_{i=1}^m f_i} = \frac{3290.06}{790} = 4.16(\text{元})$$

即790家上市公司的股票价格与其平均价格10.49元平均相差4.16元。



三、定距定比数据：标准差和方差

平均差的特点：

- **优点：**不易受极端数值的影响，能综合反映全部单位标志值的实际差异程度；
- **缺点：**用绝对值的形式消除各标志值与算术平均数离差的正负值问题，不便于作数学处理和参与统计分析运算。



三、定距定比数据：标准差和方差

标准差

是总体各单位标志值与其算术平均数离差平方的算术平均数的平方根，也称均方差。与平均差基本相同，也表示各标志值对算术平均数的平均距离

(1) 简单标准差
适用于未分组资料

$$\sigma = \sqrt{\frac{\sum_{i=1}^N (X_i - \bar{X})^2}{N}}$$

(2) 加权标准差
适用于分组资料

$$\sigma = \sqrt{\frac{\sum_{i=1}^m (X_i - \bar{X})^2 f_i}{\sum_{i=1}^m f_i}}$$



三、定距定比数据：标准差和方差

【例3.19】计算某日上海证券交易所790家上市公司股票价格的标准差。

股票价格 (元)	组中值(元) X	公司数(家) f	$(X - \bar{X})^2$	$(X - \bar{X})^2 f$
5以下	2.5	35	63.84	2234.40
5-10	7.5	457	8.94	4085.63
10-20	15	268	20.34	5451.15
20以上	25	30	210.54	6316.20
合计	—	790	—	18087.38



三、定距定比数据：标准差和方差

解：

$$\bar{X} = \frac{\sum Xf}{\sum f} = \frac{2.5 \times 35 + 7.5 \times 457 + 15 \times 268 + 25 \times 30}{35 + 457 + 268 + 30} = \frac{8285}{790} = 10.49(\text{元})$$

$$\sigma = \sqrt{\frac{\sum (X - \bar{X})^2 f}{\sum f}} = \sqrt{\frac{18087.38}{790}} = 4.78(\text{元})$$

计算结果表明，790家上市公司的股票价格的标准差为4.78元。



三、定距定比数据：标准差和方差

标准差的特点：

- 与全距相比，标准差也不受极端数值的影响，能综合反映全部单位标志值对其平均数的实际差异情况；
- 用平方的方法消除各标志值与算术平均数离差的正负值问题，可方便地用于数学处理和统计分析运算；
- 标准差的计算结果稍大于平均差，对于进行抽样估计，提高保证程度具有一定的意义。



三、定距定比数据：标准差和方差

方差

是各变量值与其平均数离差平方的平均数，即标准差的平方。

(1) 简单方差

适用于未分组资料

$$\sigma^2 = \frac{\sum_{i=1}^N (X_i - \bar{X})^2}{N}$$

(2) 加权方差

适用于分组资料

$$\sigma^2 = \frac{\sum_{i=1}^m (X_i - \bar{X})^2 f_i}{\sum_{i=1}^m f_i}$$



四、离散系数

离散系数

指消除平均数和计量单位影响后的标志变异指标，其形式表现为相对数，也称为标志变异相对数指标。

• 平均差系数

$$V_{A \cdot D} = \frac{A \cdot D}{\bar{X}} \times 100\%$$

• 标准差系数

$$V_{\sigma} = \frac{\sigma}{\bar{X}} \times 100\%$$

- 用来对比不同水平的同类现象，特别是不同类现象总体平均数代表性的大小；
- 标准差系数小的总体，其平均数的代表性大；反之，亦然。



四、离散系数

【例3.20】已知上海证券市场某年度A股平均价格为 $\bar{X} = 10.49$ (元) $\sigma = 4.78$ (元)，每股平均税后利润为 $\bar{X} = 0.28$ (元)， $\sigma = 0.16$ (元)，分别计算每股平均价格和每股平均税后利润的标准差系数，并比较它们各自平均水平代表性的大小。

解：每股平均价格的标准差系数
$$V_{\sigma} = \frac{\sigma}{\bar{X}} \times 100\% = \frac{4.78}{10.49} \times 100\% = 45.56\%$$

每股平均税后利润的标准差系数
$$V_{\sigma} = \frac{\sigma}{\bar{X}} \times 100\% = \frac{0.16}{0.28} \times 100\% = 57.14\%$$

由于平均价格的标准差系数要小于平均税后利润，可以认为A股平均价格的代表性要比每股平均税后利润的代表性大。



第三节 离中趋势的描述

对于不同类型数据，应选用恰当的离散程度测度值：

- 对于定类（分类）数据，主要用异众比率来测度其离散程度。
- 对于定序（顺序）数据，虽然也可以计算异众比率，但主要是用四分位差来测度其离散程度。
- 对于定距定比（数值型）数据，虽然可以计算异众比率、四分位差、极差和平均差等，但主要是用标准差或方差来测度其离散程度。
- 当需要对不同总体或样本数据的离散程度进行比较时，则需要使用离散系数。



第四节 分布形态的描述

一、偏态

二、峰度



一、偏态

偏态 指数据分布的不对称性

偏态系数 用来测度数据偏斜的具体程度大小的统计量

计算公式：

$$SK = \frac{\sum_{i=1}^N (X_i - \bar{X})^3}{N\sigma^3} \quad (\text{未分组数据})$$

$$SK = \frac{\sum_{i=1}^m (X_i - \bar{X})^3 f_i}{\sigma^3 \sum_{i=1}^m f_i} \quad (\text{分组数据})$$



一、偏态

偏态系数

- 偏态系数为0，表明频数分布为对称形态；偏态系数小于0，则为负偏或左偏；偏态系数大于0，则为正偏或右偏。
- 偏态系数的绝对值越大，表明左偏或右偏的程度越大。当绝对值大于1时，通常被认为是高度偏态。
- 若偏态系数的取值范围在（0.5， -1）或（-1， -0.5）两个区间内，则属于中等偏态分布，偏态系数越接近0，偏斜程度就越低。



一、偏态

【例3.22】已知某月某品牌小汽车各代理商的销售情况统计表，试计算其销售量的偏态系数。

	A	B	C	D	E	F	G
1	按销售量分组 (台)	组中值 x	代理个数 f	xf	$(x-\bar{x})^2 f$	$(x-\bar{x})^3 f$	$(x-\bar{x})^4 f$
2							
3	140-150	145	4	580	6400	-256000	10240000
4	150-160	155	9	1395	8100	-243000	7290000
5	160-170	165	16	2640	6400	-128000	2560000
6	170-180	175	27	4725	2700	-27000	270000
7	180-190	185	20	3700	0	0	0
8	190-200	195	17	3315	1700	17000	170000
9	200-210	205	10	2050	4000	80000	1600000
10	210-220	215	8	1720	7200	216000	6480000
11	220-230	225	4	900	6400	256000	10240000
12	230-240	235	5	1175	12500	625000	31250000
13	合计	—	120	22200	55400	540000	70100000



一、偏态

解：

$$\bar{X} = \frac{\sum xf}{\sum f} = \frac{22200}{120} = 185 \text{ (台)}$$

$$\sigma = \sqrt{\frac{\sum (X - \bar{X})^2 f}{\sum f}} = \sqrt{\frac{55400}{120}} = 21.49 \text{ (台)}$$

$$SK = \frac{\sum_{i=1}^m (X_i - \bar{X})^3 f_i}{\sigma^3 \sum_{i=1}^m f_i} = \frac{540000}{120 \times (21.49)^3} = 0.453$$

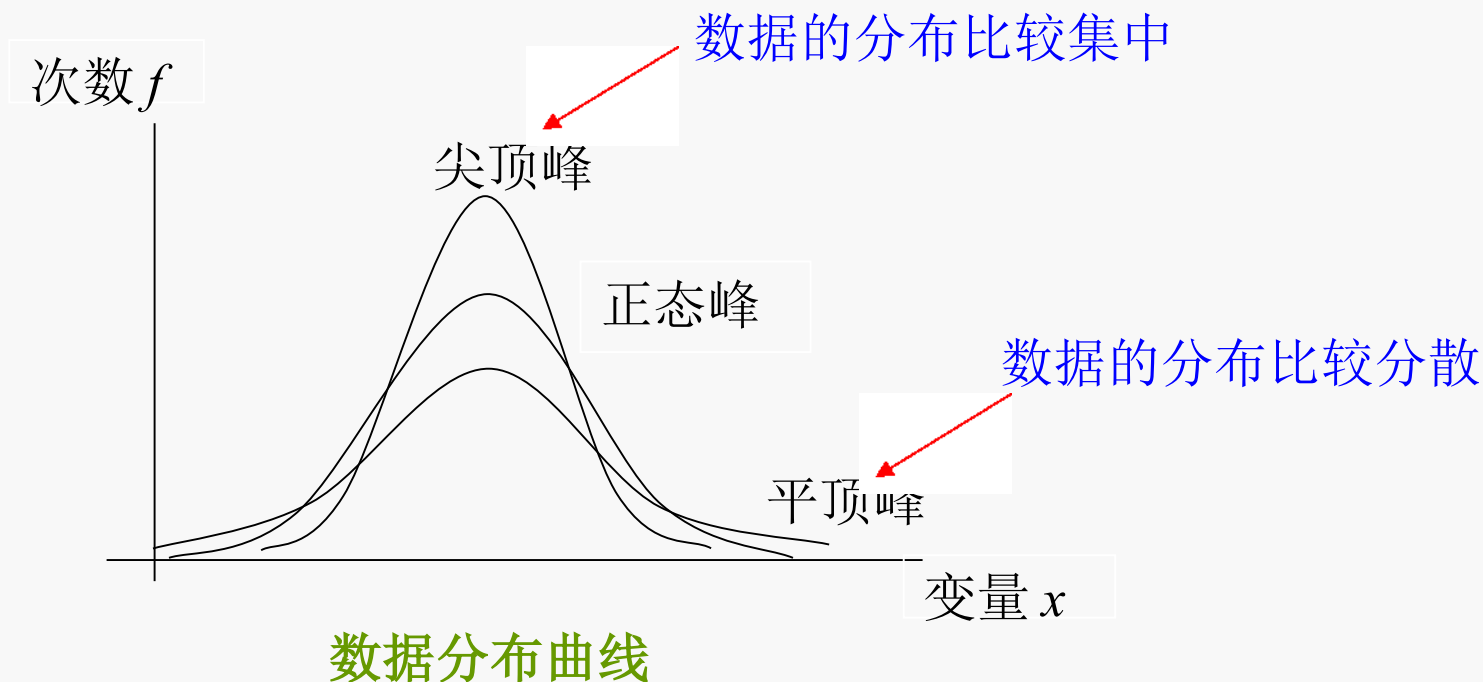
从计算结果可以看出，偏态系数为正值，但数值并不大，说明该品牌小汽车销售量呈右偏分布，但偏斜程度不大。



二、峰度

峰态

指数据分布的平峰或尖峰程度，也就是反映次数分布曲线顶端尖峭程度的综合指标。



二、峰度

峰态系数

是对数据分布峰态的度量值

$$K = \frac{\sum_{i=1}^N (X_i - \bar{X})^4}{N\sigma^4} - 3 \quad (\text{未分组数据})$$

计算公式:

$$K = \frac{\sum_{i=1}^m (X_i - \bar{X})^4 f_i}{\sigma^4 \sum_{i=1}^m f_i} - 3 \quad (\text{分组数据})$$



二、峰度

关于峰态系数，数理统计证明：

- 当 $K=0$ 时，次数分布曲线为**正态分布**
- 当 $K>0$ 时，为**尖顶峰态**
- 当 $K<0$ 时，为**平顶峰态**
- 当 $K=-1.2$ 时，次数分布曲线成为一条水平线，即各组包括的次数相同，分布形态为**矩形分布**
- 当 $K<-1.8$ 时，次数分布曲线则成为**U形分布**



二、峰度

【例3.23】仍以例3.22中的资料为例，计算其峰度系数。

解：

$$K = \frac{\sum_{i=1}^m (X_i - \bar{X})^4 f_i}{\sigma^4 \sum_{i=1}^m f_i} - 3 = \frac{70100000}{120 \times (21.49)^4} = -0.26$$

由于 $K = -0.26 < 0$ ，说明该品牌小汽车销售量的分布形态与正态分布相比略为平坦。



第五节 运用SPSS进行统计量描述

利用SPSS软件实现计算描述性统计量的过程

➤ “Frequencies” 过程

Analyze → Descriptive Statistics → Frequencies

➤ “Descriptives” 过程

Analyze → Descriptive Statistics → Descriptives

➤ “Explore” 过程

Analyze → Descriptive Statistics → Explore



