第三章　描述统计：数值方法

学习目标

学习本章后，应该做到：

1. 了解数据分布特征主要应从集中趋势的度量、离散程度的度量、分布形态的度量三个方面进行测度；
2. 理解各种测度值的特点和应用原则；
3. 掌握反映数据集中趋势和离散程度的测度方法；
4. 掌握反映分布形态的偏态与峰态的含义及测度方法。

3.1 集中趋势的度量

- 均值
- 众数
- 中位数
- 分位数
- 均值、众数、中位数的关系

一、均值（mean）

（一）均值的概念

1. 集中趋势的最常用测度值
2. 一组数据的均衡点所在
3. 体现了数据的必然性特征，消除了偶然因素的影响
4. 易受极端值的影响
5. 数值均值用于数值型数据，不能用于分类数据和顺序数据

（二）均值的算法

1、简单均值（simple mean）

设一组数据为： x_1，x_2，…，x_n

总体均值

$$\mu = \frac{x_1 + x_2 + \cdots + x_N}{N} = \frac{\sum_{i=1}^{n} x_i}{N}$$

样本均值

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

应用条件

- 通常用于对未分组的数据计算，且研究总量取决于各变量值的和

简单均值实例

【例】某班级40名同学统学的考试成绩 ：

	A	B	C	D	E	F	G	H	I	J
1	64	70	89	64	56	95	98	79	88	88
2	78	89	60	78	68	79	79	95	68	60
3	78	89	99	36	75	84	78	64	78	85
4	85	79	70	84	68	75	89	75	78	75

该班40名同学统计学的平均成绩为：

$$\bar{x}=\frac{\sum x}{n}=\frac{64+70+\mathrm{L}\ \ +78+75}{40}=\frac{3089}{40}=77.23(\text{分})$$

2、加权均值（weighted mean）

设各组变量值（组中值）为：x_1，x_2，…，x_k

对应的频数为：f_1，f_2，…，f_k

总体均值

$$\mu=\frac{x_1f_1+x_2f_2+\cdots+x_Kf_K}{f_1+f_2+\cdots+f_K}=\frac{\sum_{i=1}^{K}x_if_i}{\sum_{i=1}^{K}f_i}$$

简写为：

$$\mu=\frac{\sum xf}{\sum f}$$

- 适合分组资料的计算

应用条件

样本均值 $\bar{x}=\dfrac{x_1f_1+x_2f_2+\cdots+x_kf_k}{f_1+f_2+\cdots+f_k}=\dfrac{\sum\limits_{i=1}^{k}x_if_i}{\sum\limits_{i=1}^{n}f_i}$

简写为：

$$\bar{x}=\frac{\sum xf}{\sum f}$$

当：权数为频率 $\boxed{\dfrac{f}{\sum f}}$ 时

总体均值：$\mu = \sum x \dfrac{f}{\sum f}$

样本均值：$\bar{x} = \sum x \dfrac{f}{\sum f}$

加权均值(权数：f ）实例

某班级40名同学统计学的考试成绩：

	A	B	C	D
1	成绩（分）	频数 f_i	组中值（x_i）	$x_i f_i$
2	60以下	2	55	110
3	60～70	8	65	520
4	70～80	16	75	1200
5	80～90	10	85	850
6	90～100	4	95	380
7	合 计	40	—	3060

该班40名同学统计学的平均成绩为：

$$\mu = \frac{\sum xf}{\sum f} = \frac{3060}{40} = 76.5(\text{分})$$

加权均值(权数：$\frac{f}{\sum f}$ ）实例

	A	B	C	D
1 2 3	成绩分组（分）	频率 $\frac{f}{\sum f}$	组中值 x	$x\frac{f}{\sum f}$
4	60分以下	0.05	55	2.75
5	60～70	0.2	65	13
6	70～80	0.4	75	30
7	80～90	0.25	85	21.25
8	90分以上	0.1	95	9.5
9	合　计	1	—	76.5

$$\mu=\sum x\frac{f}{\sum f}=55\times0.05+65\times0.2+75\times0.4+85\times0.25+95\times0.1=76.5(\text{分})$$

加权均值

权数对均值的影响

甲乙两组各有10名学生，他们的考试成绩及其分布数据如下：

甲组：

考试成绩（*x*）：	0	20	100
人数分布（*f*）：	1	1	8

乙组：

考试成绩（*x*）：	0	20	100
人数分布（*f*）：	8	1	1

$$\bar{x}_{甲}=\frac{\sum xf}{\sum f}=\frac{0\times1+20\times1+100\times8}{10}=82(分)$$

$$\bar{x}_{乙}=\frac{\sum xf}{\sum f}=\frac{0\times8+20\times1+100\times1}{10}=12(分)$$

3.相对数的算术平均数

【例】某公司所属11个企业资金利润率分组资料如下表，要求计算该公司11个企业的平均资金利润率：

	A	B	C	D
1 2	资金利润率（x)%	企业数n	资金总额（万元）f	利润总额（万元）xf
3	6	4	40	2.4
4	10	3	90	9
5	15	4	140	21
6	合　计	11	270	32.4

该公司11个企业的平均资金利润率为：

$$\mu=\frac{\sum xf}{\sum f}=\frac{6\%\times40+10\%\times80+15\%\times140}{40+90+140}=\frac{32.4}{270}=12\%$$

- 权数的选择必须符合该相对数本身的计算公式。
- 权数通常为该相对数的分母指标。

4.均值的数学性质

1. 数值观测值与均值的离差之和等于0

$$\sum(x-\bar{x})=0 \quad 或 \quad \sum(x-\bar{x})f=0$$

2. 数值观测值与均值的离差平方和最小

$$\sum(x-\bar{x})^2=\min(最小) \quad 或 \quad \sum(x-\bar{x})^2 f=\min(最小)$$

3.均值易受极端值的影响

（二）调和平均数（Harmonic mean）

- 调和平均数也称为倒数平均数。
- 各变量值的倒数（$1/x_i$）的算术平均数的倒数.
- 其计算公式为：

$$\bar{x}_H=\frac{1}{\dfrac{\dfrac{1}{x_1}m_1+\dfrac{1}{x_2}m_2+\ldots+\dfrac{1}{x_n}m_n}{m_1+m_2+\ldots+m_n}}=\frac{m_1+m_2+\ldots+m_n}{\dfrac{m_1}{x_1}+\dfrac{m_2}{x_2}+\ldots+\dfrac{m_n}{x_n}}=\frac{\sum_{i=1}^{n}m_i}{\sum_{i=1}^{n}\dfrac{m_i}{x_i}}$$

（续）

- 社会经济统计中所应用的调和平均数通常是加权算术平均数的变形.
  - 已知各组变量值 x_i 和（x_i f_i），而缺乏 f_i 时，加权算术平均数通常可变形为调和平均数形式来计算。
- 【例3-3】解：

表 3-3

企 业	流通费用率（%）	商品销售额（万元）	流通费用（万元）
甲	16	1600	256
乙	10	4750	475
丙	12	4000	480
合 计	11.7	10350	1211

$$\bar{x}=\frac{\sum_{i=1}^{k}x_i f_i}{\sum_{i=1}^{k}f_i}=\frac{\sum_{i=1}^{k}(x_i f_i)}{\sum_{i=1}^{k}\frac{(x_i f_i)}{x_i}}=\frac{256+475+480}{\frac{256}{16\%}+\frac{475}{10\%}+\frac{480}{12\%}}=\frac{1211}{10350}\times 100\%=11.7\%$$

（三）几何平均数（Geometric mean）

- 几何平均数— n个变量值连乘积的n次方根。
  - 简单几何平均数
  - 加权几何平均数

$$\bar{x}_G = \sqrt[n]{x_1 \times x_2 \times \ldots \times x_n} = \sqrt[n]{\prod_{i=1}^{n} x_i}$$

$$\bar{x}_G = \sqrt[(f_1+f_2+\ldots+f_k)]{x_1^{f_1} \times x_2^{f_2} \times \ldots \times x_k^{f_k}} = \sqrt[\sum_{i=1}^{n} f_i]{\prod_{i=1}^{n} x_i^{f_i}}$$

- 适用于各个变量值之间存在连乘积关系的场合。
  - 主要用于计算现象的平均发展速度，
  - 也适用于对某些具有环比性质的比率求平均.

【例3-4】

- 某企业产品的加工要顺次经过前后衔接的五道工序。本月该企业各加工工序的合格率分别为88％、85％、90％、92％、96％，试求这五道工序的平均合格率。
- 解：本例中各工序的合格率具有环比的性质，企业产品的总合格率等于各工序合格率之连乘积。所以，所求的平均合格率应为：

$$\bar{x}_G = \sqrt[5]{88\% \times 85\% \times 90\% \times 92\% \times 96\%} = 90.31\%$$

二、众数

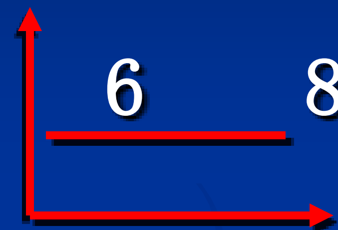
众数(mode)

1. 一组数据中出现次数最多的变量值
2. 适合于数据量较多且具有明显集中趋势时使用
3. 不受极端值的影响
4. 一组数据可能没有众数或有几个众数
5. 主要用于分类数据，也可用于顺序数据和数值型数据

众数
(不惟一性)

1. 无众数

原始数据: 10 5 9 12 6 8

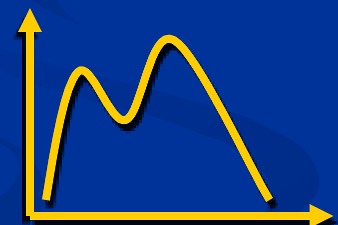


2. 一个众数

原始数据: 6 5 9 8 5 5

3. 多于一个众数

原始数据: 25 28 28 36 42 42



分组数据的众数
(要点及计算公式)

1.先找到众数组。

- 在等距数列中，众数组就是次数最多的组；
    - 在异距数列中，众数组应是频数密度*最大的组

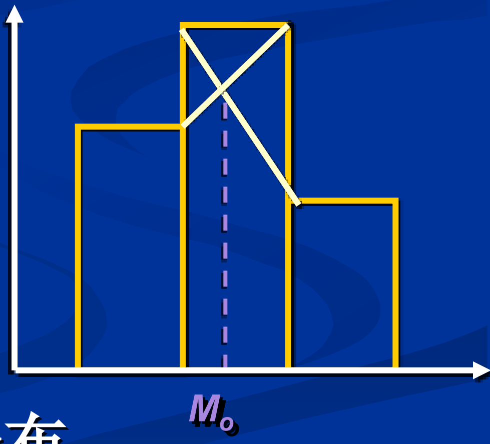
2. 众数的值与相邻两组频数的分布有关

3. 相邻两组的**频数相等**时，众数组的**组中值即为众数**

4.相邻两组的**频数不相等**时，众数采用下面的**计算公式**：

$$M_o = L + \frac{\Delta_1}{\Delta_1 + \Delta_2} \times d$$

$$M_o = U - \frac{\Delta_2}{\Delta_1 + \Delta_2} \times d$$



5.该公式假定众数组的频数在众数组内均匀分布

分组数据的众数实例

【例】某地区120家企业按利润额进行分组，结果如下：

	A	B
1 2	按利润额分组（万元）	企业数（个）
3	200～300	19
4	300～400	30
5	400～500	42
6	500～600	18
7	600以上	11
8	合　计	120

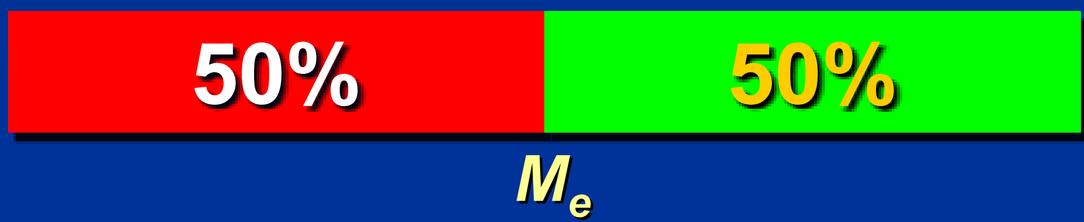
根据频数知：众数在400~500组

$$M_o = 400 + \frac{42-30}{(42-30)+(42-18)} \times 100 = 433.33(\text{万元})$$

三、中位数

中位数（median）的概念

1. 排序后处于中间位置上的变量值



2.不受极端值的影响

3.主要用于顺序数据，也可用数值型数据，但不能用于分类数据

中位数的位置确定

原始数据： $$中位数位置=\frac{n+1}{2}$$

组距分组数据： $$中位数位置=\frac{n}{2}=\frac{\sum f}{2}$$

未分组数据的中位数
(计算公式)

$$M_e = \begin{cases} X_{\left(\frac{n+1}{2}\right)} & \text{当n为奇数时} \\ \frac{1}{2}\left(X_{\frac{n}{2}} + X_{\frac{n}{2}+1}\right) & \text{当n为偶数时} \end{cases}$$

未分组数据中位数的求法 (9个数据的算例)

【例】 9个家庭的人均月收入数据

原始数据: 1500 750 780 1080 850 960 2000 1250 1630

排 序: 750 780 850 960 1080 1250 1500 1630 2000

位 置: 1 2 3 4 5 6 7 8 9

$$\text{位置} = \frac{n+1}{2} = \frac{9+1}{2} = 5$$

$$\text{中位数} = 1080$$

未分组数据中位数的求法
(10个数据的算例)

【例】：10个家庭的人均月收入数据

排 序: 660 750 780 850 960 1080 1250 1500 1630 2000

位 置: 1 2 3 4 5 6 7 8 9 10

$$位置=\frac{n+1}{2}=\frac{10+1}{2}=5.5$$

$$中位数=\frac{960+1080}{2}=1020$$

分组数据的中位数
(要点及计算公式)

1. 用于数值型分组数据
2. 根据位置公式确定中位数所在的组
3. 下限与上限计算公式分别为：

$$M_e = L + \frac{\frac{\sum f}{2} - S_{m-1}}{f_m} \times d \qquad M_e = U - \frac{\frac{\sum f}{2} - S_{m+1}}{f_m} \times d$$

4.该公式假定频数在中位数组内均匀分布

分组数据的中位数

(例题分析)

【例】某地区120家企业按利润额进行分组，结果如下:

	A	B	C	D
1	按利润额分组（万元）	企业数（个）	向上累计频数	向下累计频数
2	x	f	S_{m-1}	S_{m+1}
3	200～300	19	19	120
4	300～400	30	49	101
5	400～500	42	91	71
6	500～600	18	109	29
7	600以上	11	120	11
8	合　计	120	—	—

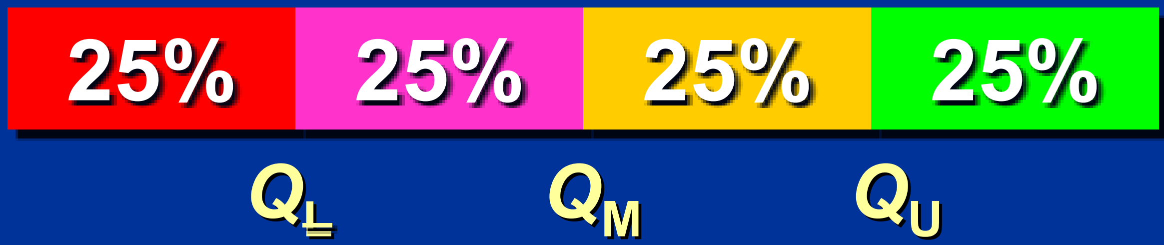
$\frac{\sum f}{2}=\frac{120}{2}=60$ 　根据向上累计频数知，中位数在400~500这一组

$$M_e=400+\frac{\frac{120}{2}-49}{42}\times 100=426.19(\text{元})$$

四、四分位数

四分位数（quartile）概念

1. 排序后处于25%和75%位置上的值



2. 不受极端值的影响