

# 第四章 数据的概括性度量

李德山

四川师范大学商学院

2022 年 3 月 18 日

# Contents

- ① 集中趋势的度量
- ② 离散趋势的度量
- ③ 偏态与峰度的度量

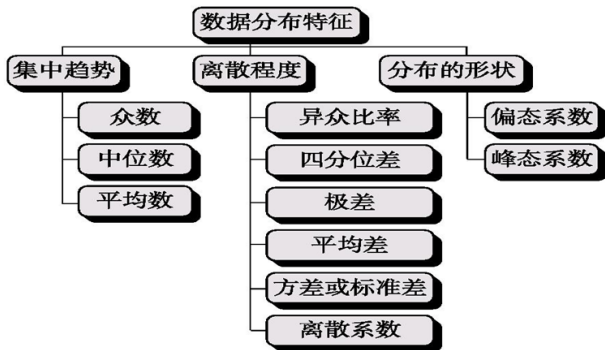
# 数据的描述

- 概括数据的特征: 如同给人画像一样
- 没有人能够记住那些巨大的数据中的所有数值, 但仍可以对数据形成一些印象。



# 数据的描述

- 数据描述的常见方法：集中趋势、离散程度、分布的形状



## ① 集中趋势的度量

## ② 离散趋势的度量

## ③ 偏态与峰度的度量

# 集中趋势

- 1. 一组数据向其中心值靠拢的倾向和程度
- 2. 测度集中趋势就是寻找数据水平的代表值或中心值
- 3. 不同类型的数据用不同的集中趋势测度值

# 众数 mode

- 一组数据中出现次数最多的变量值
- 适合于数据量较多时使用
- 不受极端值的影响
- 一组数据可能没有众数或有几个众数
- 主要用于分类数据，也可用于顺序数据和数值型数据

## 众数 mode

- 一组数据中出现次数最多的变量值
- 适合于数据量较多时使用
- 不受极端值的影响
- 一组数据可能没有众数或有几个众数
- 主要用于分类数据，也可用于顺序数据和数值型数据



## 众数 mode

- 一组数据中出现次数最多的变量值
- 适合于数据量较多时使用
- 不受极端值的影响
- 一组数据可能没有众数或有几个众数
- 主要用于分类数据，也可用于顺序数据和数值型数据

## 众数 mode

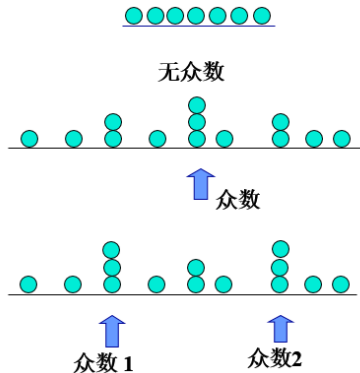
- 一组数据中出现次数最多的变量值
- 适合于数据量较多时使用
- 不受极端值的影响
- 一组数据可能没有众数或有几个众数
- 主要用于分类数据，也可用于顺序数据和数值型数据

## 众数 mode

- 一组数据中出现次数最多的变量值
- 适合于数据量较多时使用
- 不受极端值的影响
- 一组数据可能没有众数或有几个众数
- 主要用于分类数据，也可用于顺序数据和数值型数据

# 众数 mode

- 众数的不唯一性



# 中位数 median

- 一组数据按大小顺序排列后，处在数列中点位置的数值。
- 对一组数据是唯一的。
- 不受极端值的影响。
- 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。
- 各变量值与中位数的离差绝对值之和最小。

$$\sum_{i=1}^n |x_i - M_e| = \min$$

## 中位数 median

- 一组数据按大小顺序排列后，处在数列中点位置的数值。
- 对一组数据是唯一的。
- 不受极端值的影响。
- 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。
- 各变量值与中位数的离差绝对值之和最小。

$$\sum_{i=1}^n |x_i - M_e| = \min$$

# 中位数 median

- 一组数据按大小顺序排列后，处在数列中点位置的数值。
- 对一组数据是唯一的。
- 不受极端值的影响。
- 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。
- 各变量值与中位数的离差绝对值之和最小。

$$\sum_{i=1}^n |x_i - M_e| = \min$$

## 中位数 median

- 一组数据按大小顺序排列后，处在数列中点位置的数值。
- 对一组数据是唯一的。
- 不受极端值的影响。
- 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。
- 各变量值与中位数的离差绝对值之和最小。

$$\sum_{i=1}^n |x_i - M_e| = \min$$



# 中位数 median

所谓的中位数，是指将数据依大小顺序排列时，最中间的值。

首先，将各队得分依大小顺序排列看看。

A队

38	73	86	90	111	124
----	----	----	----	-----	-----

B队

71	84	85	89	90	103
----	----	----	----	----	-----

C队

59	70	77	88	95	229
----	----	----	----	----	-----

数据的个数为奇数

-1041.6	-39.0	-5.7	60.4	77.3
---------	-------	------	------	------

↓  
中位数

数据的个数为偶数

-0.4	35.2	37.8	42.2	46.1	910.3
------	------	------	------	------	-------

↑  
这两个数的平均数为中位数

数据的个数为奇数，则正中间的数即为中位数。

但若如同本次的保龄球大赛一般，数据的个数为偶数时，则第三和第四顺位的数字之平均数成为中位数。

那么，我们来算算看C队的中位数吧！

算！

$$\frac{77+88}{2} = 82.5$$

答对了！

# 中位数 median

- 计算中位数

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

1, 2, 5, 9, 11

中位数  
=5

1, 2, 5, 9, 11, 18

中位数 = (5+9) / 2 = 7

# 中位数 median

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



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

$$\text{中位数} = 1080$$

# 中位数 median

排	序:	660	750	780	850	960	1080	1250	1500	1630	2000
位	置:	1	2	3	4	5	6	7	8	9	10

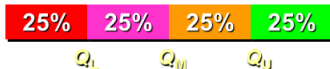


$$\text{位置} = \frac{n+1}{2} = \frac{10+1}{2} = 5.5$$

$$\text{中位数} = \frac{960+1080}{2} = 1020$$

## 四分位数 quartile

- 1. 排序后处于 25% 和 75% 位置上的值



- 2. 不受极端值的影响
- 3. 位置计算公式:

$$\begin{cases} Q_L \text{位置} = \frac{n}{4} \\ Q_U \text{位置} = \frac{3n}{4} \end{cases}$$

## 四分位数 quartile

- 不能整除时需加权平均

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
		↑					↑		

$$Q_L \text{位置} = \frac{9}{4} = 2.25$$

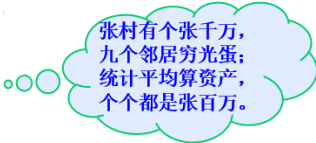
$$Q_U \text{位置} = \frac{3 \times 9}{4} = 6.75$$

$$Q_L = 780 + (850 - 780) \times 0.25 = 797.5$$

$$Q_U = 1250 + (1500 - 1250) \times 0.75 = 1437.5$$

## 平均数 mean

- 也称为均值。集中趋势的最常用测度值。
- 一组数据的均衡点所在。一组数只有一个均值。
- 各变量值与均值的离差之和等于零:  $\sum (x - \bar{x}) = 0$
- 易受极端值的影响。
- 根据总体数据计算的, 称为平均数, 记为  $\mu$ ; 根据样本数据计算的, 称为样本平均数, 记为  $\bar{x}$ 。



张村有个张千万,  
九个邻居穷光蛋;  
统计平均算资产,  
个个都是张百万。

## 简单平均数 Simple mean

- 设一组数据为： $x_1, x_2, \dots, x_n$  (总体数据  $x_N$ )
- 样本平均数:

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



## 加权平均数 Weighted mean

- 设各组的组中值为： $M_1, M_2, \dots, M_k$
- 相应的频数为： $f_1, f_2, \dots, f_k$
- 样本加权平均数：

$$\bar{x} = \frac{M_1 f_1 + M_2 f_2 + \dots + M_k f_k}{f_1 + f_2 + \dots + f_k} = \frac{\sum_{i=1}^k M_i f_i}{n}$$

# 加权平均数 Weighted mean

按师生分组	身高x(厘米)	人数f(人)
老师	226	1
学生	170	49
合 计	-	50

问：师生的平均身高为多少？

$$\bar{x} = \frac{\sum x_i}{n} = \frac{226 + 170}{2} = 198 \quad \times$$

$$\bar{x} = \frac{\sum x_i f_i}{\sum f_i} = \frac{226 + 170 \times 49}{50} = 171.12 \quad \checkmark$$

# 加权平均数 Weighted mean

某电脑公司销售量数据分组表			
按销售量分组	组中值( $M_i$ )	频数( $f_i$ )	$M_i f_i$
140~150	145	4	580
150~160	155	9	1395
160~170	165	16	2640
170~180	175	27	4725
180~190	185	20	3700
190~200	195	17	3315
200~210	205	10	2050
210~220	215	8	1720
220~230	225	4	900
230~240	235	5	1175
合计	—	120	22200

- 样本加权平均数:

$$\bar{x} = \frac{\sum_{i=1}^k M_i f_i}{n} = \frac{22200}{120} = 185$$

# 加权平均数 Weighted mean

某电脑公司销售量数据分组表			
按销售量分组	组中值( $M_i$ )	频数( $f_i$ )	$M_i f_i$
140~150	145	4	580
150~160	155	9	1395
160~170	165	16	2640
170~180	175	27	4725
180~190	185	20	3700
190~200	195	17	3315
200~210	205	10	2050
210~220	215	8	1720
220~230	225	4	900
230~240	235	5	1175
合计	—	120	22200

- 样本加权平均数:

$$\bar{x} = \frac{\sum_{i=1}^k M_i f_i}{n} = \frac{22200}{120} = 185$$

# 几何平均数

- $n$  个变量值乘积的  $n$  次方根
- 适用于对比率数据的平均
- 主要用于计算平均增长率
- 计算公式为

$$G_m = \sqrt[n]{x_1 \times x_2 \times \cdots \times x_n} = \sqrt[n]{\prod_{i=1}^n x_i}$$

- 可以看做是平均数的一种变形

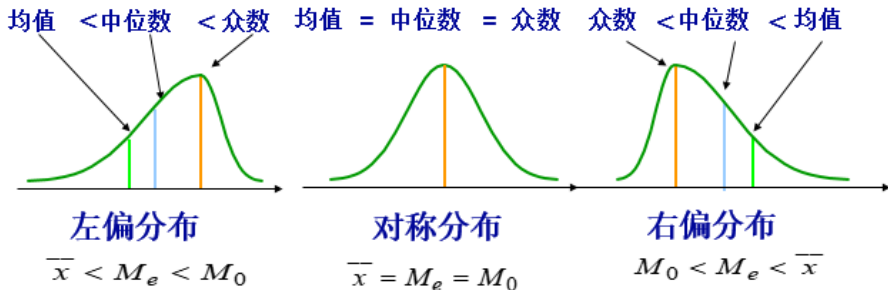
$$\lg G_m = \frac{1}{n}(\lg x_1 + \lg x_2 + \cdots + \lg x_n) = \frac{\sum_{i=1}^n \lg x_i}{n}$$

## 几何平均数的含义

- 从最初水平  $a_0$  出发，每期按平均发展速度发展，经过  $n$  期后将达到最末期水平  $a_n$
- 只与序列的最初观察值  $a_0$  和最末观察值  $a_n$  有关

$$\bar{x} = \sqrt[n]{\frac{a_1}{a_0} \times \frac{a_2}{a_1} \times \cdots \times \frac{a_n}{a_{n-1}}} = \sqrt[n]{\frac{a_n}{a_0}}$$

# 众数、中位数和算术平均数的关系



# 众数、中位数和算术平均数的特点

- 算术平均数:

- 易受极端值影响 (使用了全部数据)
- 数学性质优良, 主要用于数值型数据
- 数据对称分布或接近对称分布时应用

- 中位数:

- 不受极端值影响
- 数据分布偏斜程度较大时应用; 主要用于顺序数据

- 众数:

- 不受极端值影响
- 不具有惟一性
- 数据分布偏斜程度较大时应用; 主要用于分类数据



# 众数、中位数和算术平均数的特点

- 算术平均数：
  - 易受极端值影响 (使用了全部数据)
  - 数学性质优良, 主要用于数值型数据
  - 数据对称分布或接近对称分布时应用
- 中位数：
  - 不受极端值影响
  - 数据分布偏斜程度较大时应用; 主要用于顺序数据
- 众数：
  - 不受极端值影响
  - 不具有惟一性
  - 数据分布偏斜程度较大时应用; 主要用于分类数据

# 众数、中位数和算术平均数的特点

- 算术平均数：
  - 易受极端值影响 (使用了全部数据)
  - 数学性质优良, 主要用于数值型数据
  - 数据对称分布或接近对称分布时应用
- 中位数：
  - 不受极端值影响
  - 数据分布偏斜程度较大时应用; 主要用于顺序数据
- 众数：
  - 不受极端值影响
  - 不具有惟一性
  - 数据分布偏斜程度较大时应用; 主要用于分类数据

# 众数、中位数和算术平均数的特点

- 算术平均数：
  - 易受极端值影响 (使用了全部数据)
  - 数学性质优良, 主要用于数值型数据
  - 数据对称分布或接近对称分布时应用
- 中位数：
  - 不受极端值影响
  - 数据分布偏斜程度较大时应用; 主要用于顺序数据
- 众数：
  - 不受极端值影响
  - 不具有惟一性
  - 数据分布偏斜程度较大时应用; 主要用于分类数据

# 众数、中位数和算术平均数的特点

- 算术平均数：
  - 易受极端值影响 (使用了全部数据)
  - 数学性质优良, 主要用于数值型数据
  - 数据对称分布或接近对称分布时应用
- 中位数：
  - 不受极端值影响
  - 数据分布偏斜程度较大时应用; 主要用于顺序数据
- 众数：
  - 不受极端值影响
  - 不具有惟一性
  - 数据分布偏斜程度较大时应用; 主要用于分类数据

# 众数、中位数和算术平均数的特点

- 算术平均数：
  - 易受极端值影响 (使用了全部数据)
  - 数学性质优良, 主要用于数值型数据
  - 数据对称分布或接近对称分布时应用
- 中位数：
  - 不受极端值影响
  - 数据分布偏斜程度较大时应用; 主要用于顺序数据
- 众数：
  - 不受极端值影响
  - 不具有惟一性
  - 数据分布偏斜程度较大时应用; 主要用于分类数据

# 众数、中位数和算术平均数的特点

- 算术平均数：
  - 易受极端值影响 (使用了全部数据)
  - 数学性质优良, 主要用于数值型数据
  - 数据对称分布或接近对称分布时应用
- 中位数：
  - 不受极端值影响
  - 数据分布偏斜程度较大时应用; 主要用于顺序数据
- 众数：
  - 不受极端值影响
  - 不具有惟一性
  - 数据分布偏斜程度较大时应用; 主要用于分类数据

# 众数、中位数和算术平均数的特点

- 算术平均数：
  - 易受极端值影响 (使用了全部数据)
  - 数学性质优良, 主要用于数值型数据
  - 数据对称分布或接近对称分布时应用
- 中位数：
  - 不受极端值影响
  - 数据分布偏斜程度较大时应用; 主要用于顺序数据
- 众数：
  - 不受极端值影响
  - 不具有惟一性
  - 数据分布偏斜程度较大时应用; 主要用于分类数据

# 众数、中位数和算术平均数的特点

- 算术平均数：
  - 易受极端值影响 (使用了全部数据)
  - 数学性质优良, 主要用于数值型数据
  - 数据对称分布或接近对称分布时应用
- 中位数：
  - 不受极端值影响
  - 数据分布偏斜程度较大时应用; 主要用于顺序数据
- 众数：
  - 不受极端值影响
  - 不具有惟一性
  - 数据分布偏斜程度较大时应用; 主要用于分类数据



# 众数、中位数和算术平均数的特点

- 算术平均数：
  - 易受极端值影响 (使用了全部数据)
  - 数学性质优良, 主要用于数值型数据
  - 数据对称分布或接近对称分布时应用
- 中位数：
  - 不受极端值影响
  - 数据分布偏斜程度较大时应用; 主要用于顺序数据
- 众数：
  - 不受极端值影响
  - 不具有惟一性
  - 数据分布偏斜程度较大时应用; 主要用于分类数据

# 众数、中位数和算术平均数的特点

- 算术平均数：
  - 易受极端值影响 (使用了全部数据)
  - 数学性质优良, 主要用于数值型数据
  - 数据对称分布或接近对称分布时应用
- 中位数：
  - 不受极端值影响
  - 数据分布偏斜程度较大时应用; 主要用于顺序数据
- 众数：
  - 不受极端值影响
  - 不具有惟一性
  - 数据分布偏斜程度较大时应用; 主要用于分类数据

- ① 集中趋势的度量
- ② 离散趋势的度量
- ③ 偏态与峰度的度量

# 离散趋势的度量

- 反映各变量值远离其中心值的程度 (离散程度)
- 从另一个侧面说明了集中趋势测度值的代表程度
- 不同类型的数据有不同的离散程度测度指标

# 异众比率

- 对分类数据离散程度的测度
- 非众数组的频数占总频数的比例
- 计算公式为

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

- 用于衡量众数的代表性

# 异众比率

不同品牌饮料的频数分布			
饮料品牌	频数	比例	百分比(%)
果汁	6	0.12	12
矿泉水	10	0.20	20
绿茶	11	0.22	22
其他	8	0.16	16
碳酸饮料	15	0.30	30
合计	50	1	100

- 解:

$$\nu_r = \frac{50 - 15}{50} = 70\%$$

- 购买其他品牌饮料的人数占 70%，异众比率比较大。因此，用“碳酸饮料”代表消费者购买饮料品牌的状况，其代表性不是很好。

# 异众比率

不同品牌饮料的频数分布			
饮料品牌	频数	比例	百分比(%)
果汁	6	0.12	12
矿泉水	10	0.20	20
绿茶	11	0.22	22
其他	8	0.16	16
碳酸饮料	15	0.30	30
合计	50	1	100

- 解:

$$\nu_r = \frac{50 - 15}{50} = 70\%$$

- 购买其他品牌饮料的人数占 70%，异众比率比较大。因此，用“碳酸饮料”代表消费者购买饮料品牌的状况，其代表性不是很好。

## 四分位差 quartile deviation

- 对顺序数据离散程度的测度
- 也称为内距或四分间距
- 上四分位数与下四分位数之差

$$Q_d = Q_U - Q_L$$

- 反映了中间 50% 数据的离散程度
- 不受极端值的影响
- 用于衡量中位数的代表性



# 四分位差 quartile deviation

甲城市家庭对住房状况评价的频数分布		
回答类别	甲城市	
	户数 (户)	累计频数
非常不满意	24	24
不满意	108	132
一般	93	225
满意	45	270
非常满意	30	300
合计	300	—

- 解: 设非常不满意为 1, 不满意为 2, 一般为 3, 满意为 4, 非常满意为 5。已知  $Q_L = 2$ ,  $Q_U = 3$ , 则

$$Q_d = Q_U - Q_L = 3 - 2 = 1$$

# 四分位差 quartile deviation

甲城市家庭对住房状况评价的频数分布		
回答类别	甲城市	
	户数 (户)	累计频数
非常不满意	24	24
不满意	108	132
一般	93	225
满意	45	270
非常满意	30	300
合计	300	—

- 解: 设非常不满意为 1, 不满意为 2, 一般为 3, 满意为 4, 非常满意为 5。已知  $Q_L = 2$ ,  $Q_U = 3$ , 则

$$Q_d = Q_U - Q_L = 3 - 2 = 1$$

# 极差 range

- 一组数据的最大值与最小值之差
- 离散程度的最简单测度值
- 易受极端值影响
- 未考虑数据的分布
- 计算公式为

$$R = \max(x_i) - \min(x_i)$$

## 平均差 mean deviation

- 各变量值与其平均数离差绝对值的平均数
- 能全面反映一组数据的离散程度
- 数学性质较差，实际中应用较少
- 未分组数据：

$$M_d = \frac{\sum_{i=1}^n |x_i - \bar{x}|}{n}$$

- 组距分组数据：

$$M_d = \frac{\sum_{i=1}^k |M_i - \bar{x}| f_i}{n}$$

# 平均差 mean deviation

某电脑公司销售量数据平均差计算表				
按销售量分组	组中值( $M_i$ )	频数( $f_i$ )		
140~150	145	4	40	160
150 ~ 160	155	9	30	270
160 ~ 170	165	16	20	320
170 ~ 180	175	27	10	270
180 ~ 190	185	20	0	0
190 ~ 200	195	17	10	170
200 ~ 210	205	10	20	200
210 ~ 220	215	8	30	240
220 ~ 230	225	4	40	160
230 ~ 240	235	5	50	250
合计	—	120	—	2040

• 解：

$$M_d = \frac{\sum_{i=1}^k |M_i - \bar{x}| f_i}{n} = \frac{2040}{120} = 17$$

• 每一天的销售量与平均数相比，平均相差 17 台。

# 平均差 mean deviation

某电脑公司销售量数据平均差计算表				
按销售量分组	组中值( $M_i$ )	频数( $f_i$ )		
140~150	145	4	40	160
150 ~ 160	155	9	30	270
160 ~ 170	165	16	20	320
170 ~ 180	175	27	10	270
180 ~ 190	185	20	0	0
190 ~ 200	195	17	10	170
200 ~ 210	205	10	20	200
210 ~ 220	215	8	30	240
220 ~ 230	225	4	40	160
230 ~ 240	235	5	50	250
合计	—	120	—	2040

• 解：

$$M_d = \frac{\sum_{i=1}^k |M_i - \bar{x}| f_i}{n} = \frac{2040}{120} = 17$$

• 每一天的销售量与平均数相比，平均相差 17 台。

# 方差和标准差 variance and standard deviation

- 数据离散程度的最常用测度值
- 反映了各变量值与均值的平均差异
- 根据总体数据计算的，称为总体方差  $\sigma^2$  或总体标准差  $\sigma$ ；根据样本数据计算的，称为样本方差  $s^2$  或样本标准差  $s$

# 总体方差和样本方差

	总体方差	样本方差
未分组数据	$\sigma^2 = \frac{\sum_{i=1}^N (X_i - \bar{X})^2}{N}$	$s^2 = \frac{\sum_{i=1}^n (x_i - \bar{x})^2}{n-1}$
分组数据	$\sigma^2 = \frac{\sum_{i=1}^K (X_i - \bar{X})^2 f_i}{\sum_{i=1}^K f_i}$	$s^2 = \frac{\sum_{i=1}^k (x_i - \bar{x})^2 f_i}{\sum_{i=1}^k f_i - 1}$



# 样本方差和样本标准差

## 方差的计算公式

未分组数据

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

组距分组数据

$$s^2 = \frac{\sum_{i=1}^k (M_i - \bar{x})^2 f_i}{n-1}$$

## 标准差的计算公式

未分组数据

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

组距分组数据

$$s = \sqrt{\frac{\sum_{i=1}^k (M_i - \bar{x})^2 f_i}{n-1}}$$

# 自由度

- 自由度是指数据个数与附加给独立的观测值的约束或限制的个数之差
- 从字面涵义来看，自由度是指一组数据中可以自由取值的个数
- 当样本数据的个数为  $n$  时，若样本平均数确定后，则附加给  $n$  个观测值的约束个数就是 1 个，因此只有  $n - 1$  个数据可以自由取值，其中必有一个数据不能自由取值
- 按着这一逻辑，如果对  $n$  个观测值附加的约束个数为  $k$  个，自由度则为  $n - k$

# 自由度

- 自由度是指数据个数与附加给独立的观测值的约束或限制的个数之差
- 从字面涵义来看，自由度是指一组数据中可以自由取值的个数
- 当样本数据的个数为  $n$  时，若样本平均数确定后，则附加给  $n$  个观测值的约束个数就是 1 个，因此只有  $n - 1$  个数据可以自由取值，其中必有一个数据不能自由取值
- 按着这一逻辑，如果对  $n$  个观测值附加的约束个数为  $k$  个，自由度则为  $n - k$

# 自由度

- 自由度是指数据个数与附加给独立的观测值的约束或限制的个数之差
- 从字面涵义来看，自由度是指一组数据中可以自由取值的个数
- 当样本数据的个数为  $n$  时，若样本平均数确定后，则附加给  $n$  个观测值的约束个数就是 1 个，因此只有  $n - 1$  个数据可以自由取值，其中必有一个数据不能自由取值
- 按着这一逻辑，如果对  $n$  个观测值附加的约束个数为  $k$  个，自由度则为  $n - k$

# 自由度

- 自由度是指数据个数与附加给独立的观测值的约束或限制的个数之差
- 从字面涵义来看，自由度是指一组数据中可以自由取值的个数
- 当样本数据的个数为  $n$  时，若样本平均数确定后，则附加给  $n$  个观测值的约束个数就是 1 个，因此只有  $n - 1$  个数据可以自由取值，其中必有一个数据不能自由取值
- 按着这一逻辑，如果对  $n$  个观测值附加的约束个数为  $k$  个，自由度则为  $n - k$

## 自由度

- 样本有 3 个数值，即  $x_1 = 2$ ,  $x_2 = 4$ ,  $x_3 = 9$ ，则  $\bar{x} = 5$ 。当  $\bar{x} = 5$  确定后， $x_1$ ,  $x_2$  和  $x_3$  有两个数据可以自由取值，另一个则不能自由取值，比如即  $x_1 = 6$ ,  $x_2 = 7$ ，那么即  $x_3$  则必然取 2，而不能取其他值
- 为什么样本方差的自由度为什么是  $n - 1$  呢？因为在计算离差平方和时，必须先求出样本均值  $\bar{x}$ ，而  $\bar{x}$  则是附件给离差平方和的一个约束，因此，计算离差平方和时只有  $n - 1$  个独立的观测值，而不是  $n$  个
- 样本方差用自由度去除，其原因可从多方面解释，从实际应用角度看，在抽样估计中，当用样本方差  $s^2$  去估计总体方差  $\sigma^2$  时，它是  $\sigma^2$  的无偏估计量

# 自由度



# 样本标准差

某电脑公司销售量数据平均差计算表				
按销售量分组	组中值( $M_i$ )	频数( $f_i$ )		
140~150	145	4	40	160
150 ~ 160	155	9	30	270
160 ~170	165	16	20	320
170 ~180	175	27	10	270
180 ~ 190	185	20	0	0
190 ~ 200	195	17	10	170
200 ~ 210	205	10	20	200
210 ~220	215	8	30	240
220 ~230	225	4	40	160
230 ~240	235	5	50	250
合计	—	120	—	55400

$$s = \sqrt{\frac{\sum_{i=1}^k (M_i - \bar{x})^2 f_i}{n - 1}} = \sqrt{\frac{55400}{120 - 1}} = 21.58$$

含义：每一天的销售量与平均数相比，平均相差 21.58 台



# 样本标准差

某电脑公司销售量数据平均差计算表				
按销售量分组	组中值( $M_i$ )	频数( $f_i$ )		
140~150	145	4	40	160
150 ~ 160	155	9	30	270
160 ~170	165	16	20	320
170 ~180	175	27	10	270
180 ~ 190	185	20	0	0
190 ~ 200	195	17	10	170
200 ~ 210	205	10	20	200
210 ~220	215	8	30	240
220 ~230	225	4	40	160
230 ~240	235	5	50	250
合计	—	120	—	55400

$$s = \sqrt{\frac{\sum_{i=1}^k (M_i - \bar{x})^2 f_i}{n - 1}} = \sqrt{\frac{55400}{120 - 1}} = 21.58$$

含义：每一天的销售量与平均数相比，平均相差 21.58 台

# 样本标准差

- 平均值和标准差是人生的两条鞭子
  - 平均值让我们不甘落后
  - 标准差让我们不随大流
  - 年轻时百尺竿头更进一步
  - 中年后比上不足比下有余

## 标准分数 standard score

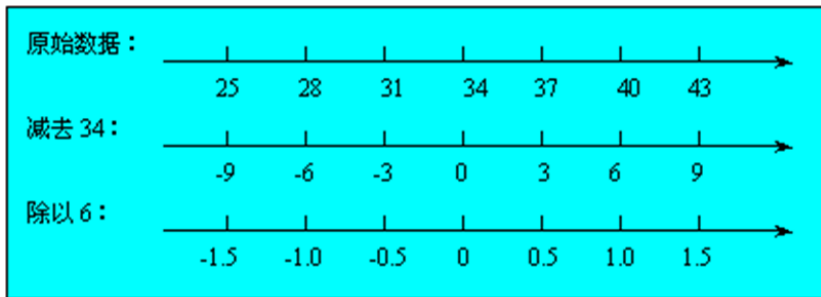
- 也称标准化值
- 对某一个值在一组数据中相对位置的度量
- 可用于判断一组数据是否有离群点 (outlier)
- 用于对变量的标准化处理

$$z_i = \frac{x_i - \bar{x}}{s}$$

- $z$  大于 0 说明观测值大于均值;  $z$  小于 0 说明观测值小于均值;  
 $z=1.2$  说明观测值比均值大 1.2 倍的标准差

## 标准分数 standard score

- $z$  分数只是将原始数据进行了线性变换，它并没有改变一个数据在该组数据中的位置，也没有改变该组数分布的形状，而只是使该组数据均值为 0，标准差为 1



# 标准分数 standard score

9个家庭人均月收入标准化值计算表		
家庭编号	人均月收入（元）	标准化值 $z$
1	1500	0.695
2	750	-1.042
3	780	-0.973
4	1080	-0.278
5	850	-0.811
6	960	-0.556
7	2000	1.853
8	1250	0.116
9	1630	0.996

# 经验法则

- 当一组数据对称分布时
- 约有 68% 的数据在平均数加减 1 个标准差的范围之内
- 约有 95% 的数据在平均数加减 2 个标准差的范围之内
- 约有 99% 的数据在平均数加减 3 个标准差的范围之内

## 切比雪夫不等式 Chebyshev's inequality

- 如果一组数据不是对称分布，经验法则就不再适用，这时可使用切比雪夫不等式，它对任何分布形状的数据都适用
- 切比雪夫不等式提供的是“下界”，也就是“所占比例至少是多少”
- 对于任意分布形态的数据，根据切比雪夫不等式，至少有  $1 - 1/k^2$  的数据落在平均数加减  $k$  个标准差之内。其中  $k$  是大于 1 的任意值，但不一定是整数

# 切比雪夫不等式 Chebyshev' s inequality

- 对于  $k=2, 3, 4$ , 该不等式的含义是
- 1. 至少有 75% 的数据落在平均数加减 2 个标准差的范围之内
- 2. 至少有 89% 的数据落在平均数加减 3 个标准差的范围之内
- 3. 至少有 94% 的数据落在平均数加减 4 个标准差的范围之内



# 离散系数 coefficient of variation

- 标准差与其相应的均值之比

$$\nu_s = \frac{s}{\bar{x}}$$

- 对数据相对离散程度的测度
- 消除了数据水平高低和计量单位的影响
- 用于对不同组别数据离散程度的比较

# 离散系数

- 试比较产品销售额与销售利润的离散程度

某管理局所属8家企业的产品销售数据		
企业编号	产品销售额（万元） $x_1$	销售利润（万元） $x_2$
1	170	8.1
2	220	12.5
3	390	18.0
4	430	22.0
5	480	26.5
6	650	40.0
7	950	64.0
8	1000	69.0

- $\nu_1 = \frac{309.19}{536.25} = 0.577, \nu_2 = \frac{23.09}{32.5215} = 0.710$

- 结论： $\nu_1 < \nu_2$  说明销售额的离散程度小于销售利润的离散程度

# 离散系数

- 试比较产品销售额与销售利润的离散程度

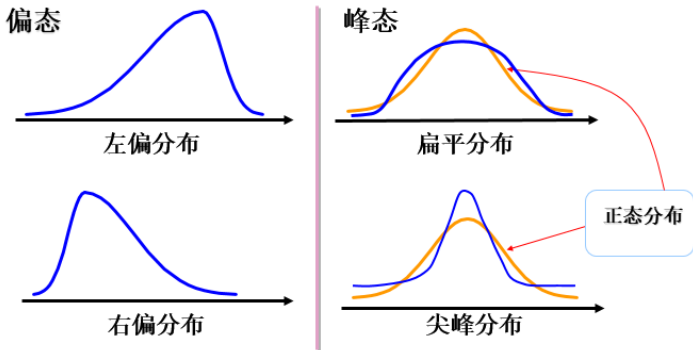
某管理局所属8家企业的产品销售数据		
企业编号	产品销售额（万元） $x_1$	销售利润（万元） $x_2$
1	170	8.1
2	220	12.5
3	390	18.0
4	430	22.0
5	480	26.5
6	650	40.0
7	950	64.0
8	1000	69.0

- $\nu_1 = \frac{309.19}{536.25} = 0.577, \nu_2 = \frac{23.09}{32.5215} = 0.710$
- 结论:  $\nu_1 < \nu_2$  说明销售额的离散程度小于销售利润的离散程度

- ① 集中趋势的度量
- ② 离散趋势的度量
- ③ 偏态与峰度的度量

# 偏态与峰态的度量

- 偏态和峰度的类型



## 偏态 skewness

- 统计学家 Pearson 于 1895 年首次提出
- 数据分布偏斜程度的测度
- 偏态系数等于 0 为对称分布
- 偏态系数大于 0 为右偏分布
- 偏态系数小于 0 为左偏分布
- 偏态系数大于 1 或小于 -1, 被称为高度偏态分布; 偏态系数越接近 0, 偏斜程度就越低

# 偏态系数

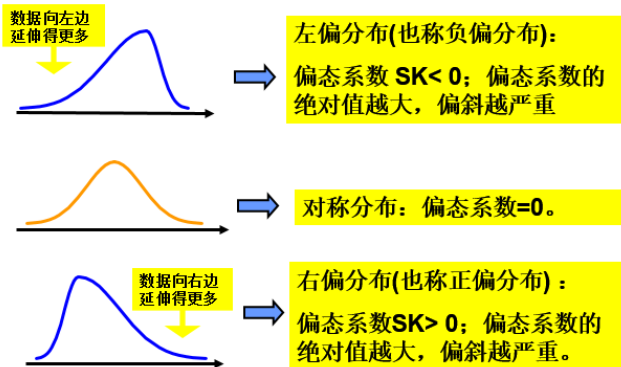
- 根据原始数据计算

$$SK = \frac{n \sum (x_i - \bar{x})^3}{(n-1)(n-2)s^3}$$

- 根据分组数据计算

$$SK = \frac{\sum_{i=1}^k (M_i - \bar{x})^3 f_i}{ns^3}$$

# 偏态系数





## 峰态 kurtosis

- 统计学家 Pearson 于 1905 年首次提出
- 数据分布扁平程度的测度
- 峰态系数等于 0 扁平峰度适中
- 峰态系数小于 0 为扁平分布
- 峰态系数大于 0 为尖峰分布

# 峰态系数

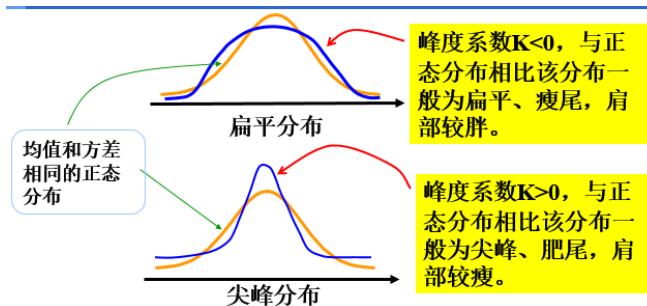
- 根据原始数据计算

$$K = \frac{n(n+1) \sum (x_i - \bar{x})^4 - 3 \left[ \sum (x_i - \bar{x})^2 \right]^2 (n-1)}{(n-1)(n-2)(n-3)s^4}$$

- 根据分组数据计算

$$K = \frac{\sum_{i=1}^k (M_i - \bar{x})^4 f_i}{ns^4} - 3$$

# 峰度系数



# 用 EXCEL 计算描述统计量步骤

- 第 1 步: 选择【工具】下拉菜单
- 第 2 步: 选择【数据分析】选项
- 第 3 步: 在分析工具中选择【描述统计】, 然后选择【确定】
- 第 4 步: 当对话框出现时
  - 在【输入区域】方框内键入数据区域
  - 在【输出选项】中选择输出区域
  - 选择【汇总统计】
  - 选择【确定】

# 用 EXCEL 计算描述统计量步骤

- 第 1 步: 选择【工具】下拉菜单
- 第 2 步: 选择【数据分析】选项
- 第 3 步: 在分析工具中选择【描述统计】, 然后选择【确定】
- 第 4 步: 当对话框出现时
  - 在【输入区域】方框内键入数据区域
  - 在【输出选项】中选择输出区域
  - 选择【汇总统计】
  - 选择【确定】

# 用 EXCEL 计算描述统计量步骤

- 第 1 步: 选择【工具】下拉菜单
- 第 2 步: 选择【数据分析】选项
- 第 3 步: 在分析工具中选择【描述统计】, 然后选择【确定】
- 第 4 步: 当对话框出现时
  - 在【输入区域】方框内键入数据区域
  - 在【输出选项】中选择输出区域
  - 选择【汇总统计】
  - 选择【确定】

# 用 EXCEL 计算描述统计量步骤

- 第 1 步: 选择【工具】下拉菜单
- 第 2 步: 选择【数据分析】选项
- 第 3 步: 在分析工具中选择【描述统计】, 然后选择【确定】
- 第 4 步: 当对话框出现时
  - 在【输入区域】方框内键入数据区域
  - 在【输出选项】中选择输出区域
  - 选择【汇总统计】
  - 选择【确定】

# 用 EXCEL 计算描述统计量步骤

- 第 1 步: 选择【工具】下拉菜单
- 第 2 步: 选择【数据分析】选项
- 第 3 步: 在分析工具中选择【描述统计】, 然后选择【确定】
- 第 4 步: 当对话框出现时
  - 在【输入区域】方框内键入数据区域
  - 在【输出选项】中选择输出区域
  - 选择【汇总统计】
  - 选择【确定】



# 用 EXCEL 计算描述统计量步骤

- 第 1 步: 选择【工具】下拉菜单
- 第 2 步: 选择【数据分析】选项
- 第 3 步: 在分析工具中选择【描述统计】, 然后选择【确定】
- 第 4 步: 当对话框出现时
  - 在【输入区域】方框内键入数据区域
  - 在【输出选项】中选择输出区域
  - 选择【汇总统计】
  - 选择【确定】

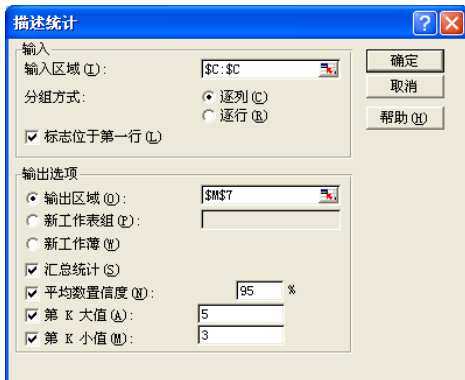
# 用 EXCEL 计算描述统计量步骤

- 第 1 步: 选择【工具】下拉菜单
- 第 2 步: 选择【数据分析】选项
- 第 3 步: 在分析工具中选择【描述统计】, 然后选择【确定】
- 第 4 步: 当对话框出现时
  - 在【输入区域】方框内键入数据区域
  - 在【输出选项】中选择输出区域
  - 选择【汇总统计】
  - 选择【确定】

# 用 EXCEL 计算描述统计量步骤

- 第 1 步: 选择【工具】下拉菜单
- 第 2 步: 选择【数据分析】选项
- 第 3 步: 在分析工具中选择【描述统计】, 然后选择【确定】
- 第 4 步: 当对话框出现时
  - 在【输入区域】方框内键入数据区域
  - 在【输出选项】中选择输出区域
  - 选择【汇总统计】
  - 选择【确定】

# 用 EXCEL 计算描述统计量步骤



# 安装步骤

- 具体安装参见《统计学 B》-课件 1.nb

## 参考资料

- 贾俊平.《统计学》(第八版) [M]. 北京: 中国人民大学出版社, 2021。
- 向蓉美等. 统计学 (第 2 版) [M]. 北京: 机械工业出版社, 2017。

Q&A

THANK YOU