

第一讲-1 数据的描述性分析



教学目的和要求

- ❖ 通过本讲的学习，学生应该熟练掌握描述数据特征，即数据的集中趋势、离散趋势的分析方法，要求学生掌握各种测度指标的含义和计算方法，并且能够运用这些指标来研究事物或现象的总体特征和变化规律。



本讲内容



❖ 描述数据特征的意义



❖ 集中趋势测度指标

种类
计算
适用情况



❖ 离散趋势测度指标

种类
计算
适用情况



❖ 位置测度指标

种类
计算
适用情况

❖ 分布形态

左偏
对称
右偏

❖ 箱索图

绘制方法
作用



描述数据特征的意义

- 1.集中趋势指标的最一般意义：作为总体的代表水平同其他同质的总体进行比较；反映的是同质总体的共性、集中性。
- 2.离散趋势指标反映的是个性和分散性，用来衡量集中趋势指标的代表性强弱。



集中趋势测度指标（平均数、中位数、众数）

❖ 掌握计算方法

❖ 掌握每种指标的适用情况



集中趋势指标-----平均数

- ❖ 衡量变量分布中心的指标
- ❖ 最常用的 集中趋势指标
- ❖ 容易受极端值的影响
 - 极端值：远离分布中心的数值



平均数的种类

- ❖ 简单算术平均数
- ❖ 加权算术平均数
- ❖ 几何平均数



简单算术平均数

❖ 公式:
$$\bar{X} = \frac{X_1 + X_2 + \cdots + X_n}{n} = \frac{\sum_{i=1}^n X_i}{n} = \frac{\sum X}{n}$$
$$= X_1 \frac{1}{n} + X_2 \frac{1}{n} + \cdots + X_n \frac{1}{n}$$

❖ 适用情况

◆ 资料未分组

◆ 每一个变量值的作用相同

❖ 影响平均数大小的因素只有变量值



加权算术平均数

❖ 定义:将各变量值分别乘以代表该变量值重要程度的权数,然后用此乘积之和除以权数之和,所得的商为加权算术平均数。

❖ 公式:

$$\begin{aligned}\bar{X} &= \frac{X_1W_1 + X_2W_2 + \cdots + X_kW_k}{W_1 + W_2 + \cdots + W_k} = \frac{\sum_{i=1}^k X_iW_i}{\sum_{i=1}^k W_i} = \frac{\sum XW}{\sum W} \\ &= X_1 \frac{W_1}{\sum W} + X_2 \frac{W_2}{\sum W} + \cdots + X_k \frac{W_k}{\sum W}\end{aligned}$$



加权平均数

(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**

$$\bar{x}_{\text{甲}} = \frac{\sum_{i=1}^n x_i}{n} = \frac{0 \times 1 + 20 \times 1 + 100 \times 8}{10} = 82(\text{分})$$

$$\bar{x}_{\text{乙}} = \frac{\sum_{i=1}^n x_i}{n} = \frac{0 \times 8 + 20 \times 1 + 100 \times 1}{10} = 12(\text{分})$$



几何平均数

公式: $\overline{X}_g = \sqrt[n]{x_1 * x_2 * \dots * x_i} = \sqrt[n]{\prod x_i}$

- ◆几何平均数适用于比例和速度等
- ◆相对数的平均计算



几何平均数的应用1

如： 某产品的生产需要四个工序，第 1、2、3 和 4 工序的产品合格率分别为 97%、93%、95% 和 91%，求平均各工序的合格率.

$$\bar{X} = \sqrt[4]{0.97 \times 0.93 \times 0.95 \times 0.91} = 94.0\%$$



几何平均数的应用2

2013年各月全国住宅价格环比指数为

时间	环比指数	时间	环比指数
1月	100.7	7月	100.8
2月	101.1	8月	100.9
3月	101.2	9月	100.7
4月	101.1	10月	100.6
5月	100.9	11月	100.6
6月	100.8	12月	100.4

平均环比速度

平均环比速度

$$\sqrt[12]{1.007 \times 1.011 \times \cdots \times 1.004} = 100.82\%$$

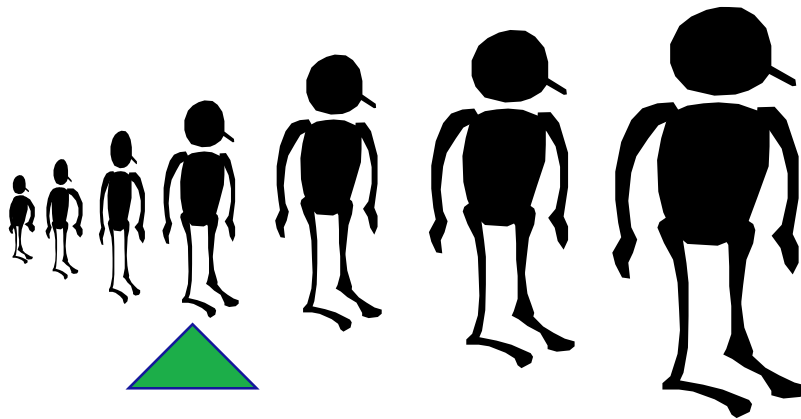


集中趋势指标2--中位数(Median)

计算方法: 将变量数列的各观察值按自小到大的顺序排列处于中间位置的数值即为中位数.

中位数所在的位置项数 = $(n+1) / 2$

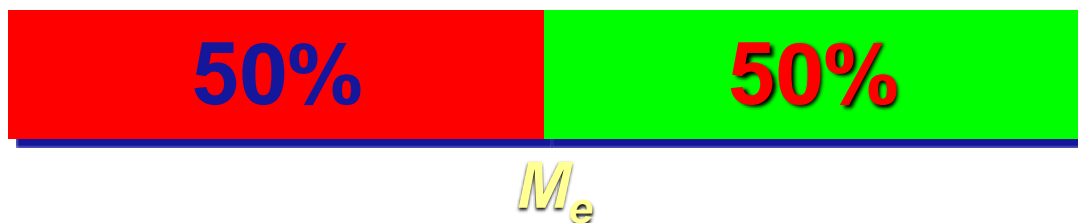
实用情况: 当数列中有极端值存在时, 采用中位数求变量值的一般水平比用算术平均数好.



中位数计算方法

❖ 对于未分组数据

□ 排序后处于中间位置上的值。不受极端值影响



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



中位数计算举例2 (5个数据的算例)

❖ 原始数据:	24	22	21	26	20
❖ 排 序:	20	21	22	24	26
❖ 位 置:	1	2	3	4	5



$$\text{位置} = \frac{N+1}{2} = \frac{5+1}{2} = 3$$


中位数 = 22



中位数计算举例 (N=6)

原始资料: **10.3 4.9 8.9 11.7 6.3 7.7**

按顺序排列: **4.9 6.3 7.7 8.9 10.3 11.7**

位置: **1 2 3  4 5 6**

中位数所在的位置为: $\frac{N+1}{2} = \frac{6+1}{2} = 3.5$

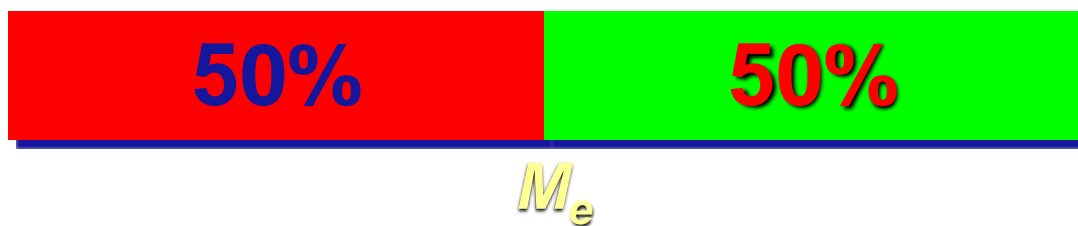
中位数 = **$(7.7 + 8.9) / 2 = 8.3$**



中位数计算方法

❖ 对于分组数据

□ 排序后处于中间位置上的值。不受极端值影响

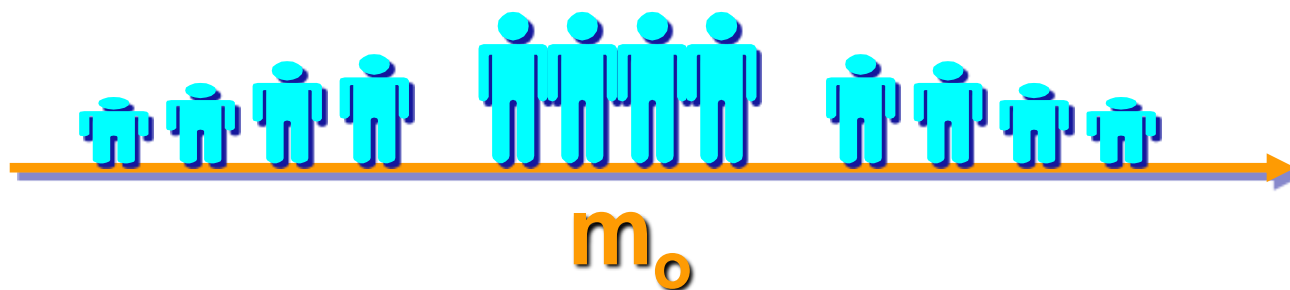


□ 确定中位数项次 $(n+1) / 2$



集中趋势指标3--众数(Mode)

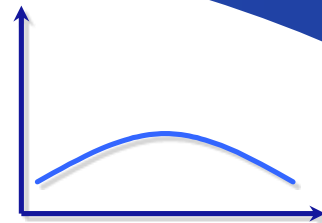
- ❖ 出现次数最多的那个变量值
- ❖ 是一个常用的集中趋势指标
- ❖ 它不受极端值的影响
- ❖ 并非所有的数列都存在众数



(众数的不唯一性)

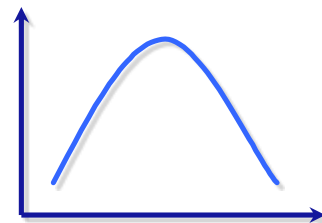
❖ 无众数

原始数据: **10 5 9 12 6 8**



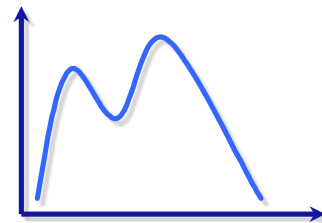
❖ 一个众数

原始数据: **6 5 9 8 5 5**



❖ 多于一个众数

原始数据: **25 28 28 36 42 42**



众数、中位数、平均数的特点和应用

1. 平均数

- 易受极端值影响
- 数学性质优良，实际中最常用
- 数据对称分布或接近对称分布时代表性较好

2. 中位数

- 不受极端值影响
- 数据分布偏斜程度较大时代表性接好

3. 众数

- 不受极端值影响
- 具有不惟一性
- 数据分布偏斜程度较大且有明显峰值时代表性较好

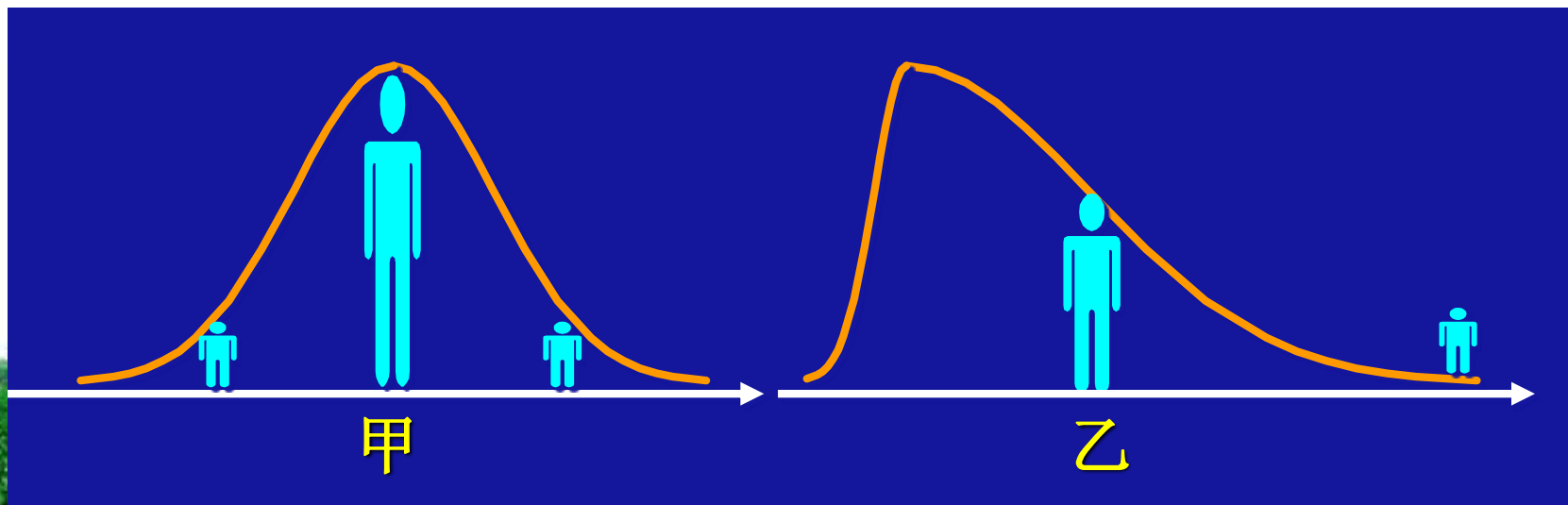


怎样评价集中趋势代表值？

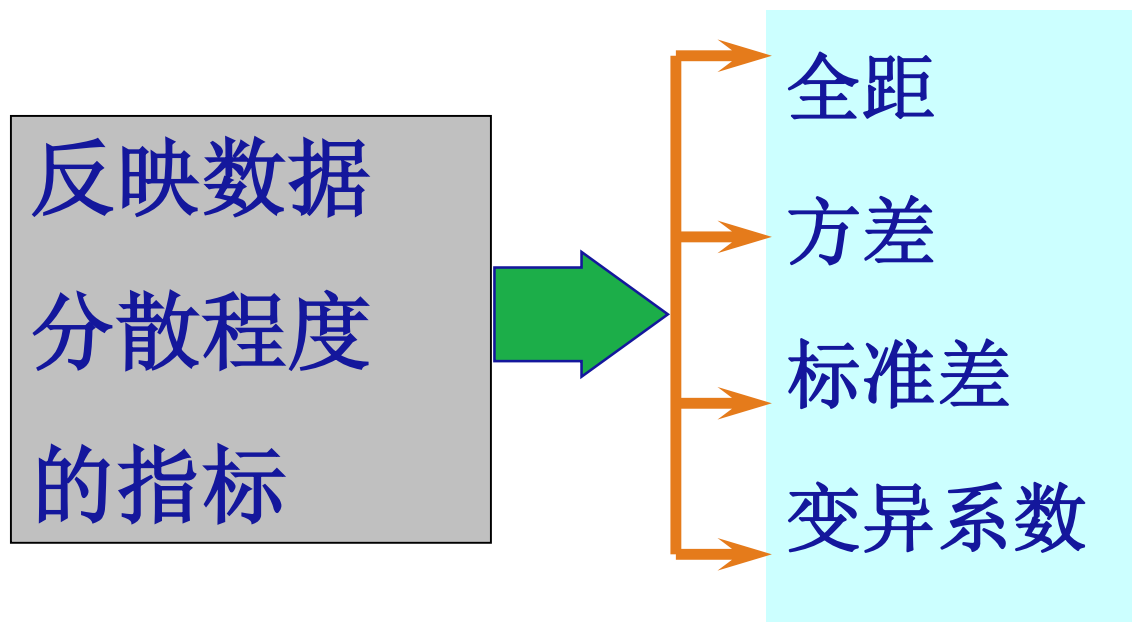
1. 假定有两个地区每人的平均收入数据，其中甲地区的平均收入为**5000**元，乙地区的平均收入为**3000**元。你如何评价两个地区的收入状况？
2. 如果平均收入的多少代表了该地区的生活水平，你能否认为甲地区的平均生活水平就高于乙地区呢？
3. 要回答这些问题，首先需要搞清楚这里的平均收入是否能代表大多数人的收入水平。如果甲地区有少数几个富翁，而大多数人的收入都很低，虽然平均收入很高，但多数人生活水平仍然很低。相反，乙地区多数人的收入水平都在**3000**元左右，虽然平均收入看上去不如甲地区，但多数人的生活水平却比甲地区高，原因是甲地区的收入差距大于乙地区。

怎样评价集中趋势代表值？

- ☺ 仅仅知道数据的集中趋势是远远不够的，还必须考虑数据之间的差距有多大。数据之间的差距用统计语言来说就是数据的离散程度。数据的离散程度越大，各描述统计量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。



离散趋势测度指标



离散程度指标

- ❖ 掌握这些指标的作用
- ❖ 掌握计算它们的方法
- ❖ 掌握每种指标的适用情况
- ❖ 掌握这些指标的优缺点



全距 (Range)

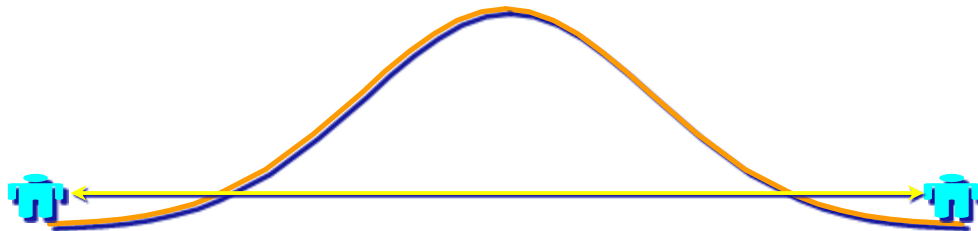
❖ 全距 = 最大值 - 最小值

❖ 原始资料: **17 16 21 18 13 16 12 11**

❖ 顺序排列: **11 12 13 16 16 17 18 21**

❖ 全距 = **21 - 11 = 10**

优缺点: 离散程度的最简单测度值;
未考虑数据的分布;
易受极端值影响;



方差和标准差 (variance and standard deviation)

1. 数据离散程度的最常用测度值
2. 反映各变量值与均值的平均差异
3. 根据总体数据计算的，称为总体方差(标准差)，记为 $\sigma^2(\sigma)$ ；根据样本数据计算的，称为样本方差(标准差)，记为 $s^2(s)$



方 差

1.总体方差 $\sigma^2 = \frac{\sum (X - \mu)^2}{N}$

2.样本方差 $S^2 = \frac{\sum (X - \bar{X})^2}{n - 1}$



样本方差计算**1**(未分组)

❖ 原始数据: **17 16 21 18 13 16 12 11**

$$S^2 = \frac{\sum (X - \bar{X})^2}{n-1}$$

$$\bar{X} = \frac{\sum X}{n} = 15.5$$

$$S^2 = \frac{(17-15.5)^2 + (16-15.5)^2 + \text{☹} + (11-15.5)^2}{8-1}$$
$$= 11.14$$



样本方差计算2续(已分组)

问题：◆11.14说明什么

◆优点:离散程度可以量化

缺点:方差计算结果会给人以夸大离散程度规模的效果，使人们不易达到直观认识离散程度的目的；方差的计量单位与原观察值得计量单位不一致。



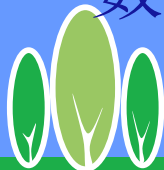
标准差(方差的平方根)

◆ 计算公式

◆展示的信息：一组数据对其均值为代表的中心的某种偏离程度。

◆优点：反映的一组数据的离散程度。标准差(或方差)较小的分布一定都是比较集中在均值附近的，反之则是比较分散的。

缺点：计算起来比较麻烦。标准差也是根据全部数据来计算的,但是它也会受到极端值的影响。



样本方差和标准差

(sample variance and standard deviation)

方差的计算公式

$$S^2 = \frac{\sum (X - \bar{X})^2}{n-1} \quad (n-1 \text{ 为样本的自由度})$$

标准差的计算公式

$$S = \sqrt{\frac{\sum (X - \bar{X})^2}{n-1}}$$



总体方差和标准差

(Population variance and Standard deviation)

方差的计算公式

$$\sigma^2 = \frac{\sum (X - \mu)^2}{N}$$

标准差的计算公式

$$\sigma = \sqrt{\frac{\sum (X - \mu)^2}{N}}$$



标准差的应用

- ❖ 标准差度量投资风险
- ❖ 标准差度量产品质量的稳定性
- ❖ 标准差度量企业的生产及服务的质量标准



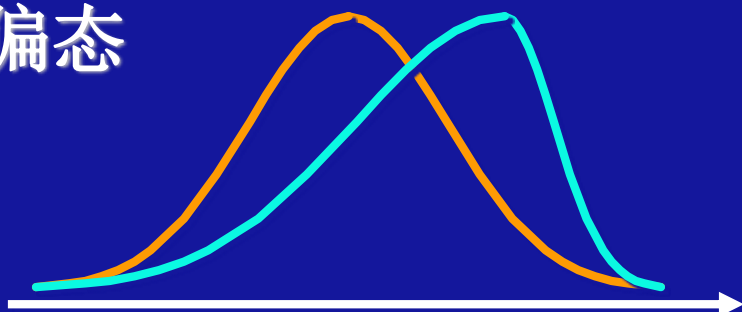
集中趋势指标与离散程度指标的关系

- 离散程度指标大，说明总体分散或者说总体中各标志值离集中趋势指标远，那么集中趋势指标代表性就小。
- 离散程度指标小，说明总体集中或者说总体中各标志值离集中趋势指标近，那么集中趋势指标代表性就大。

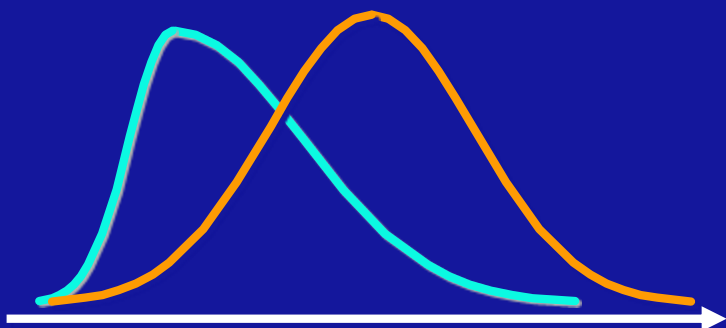


数据分布的形状—偏态与峰态

偏态

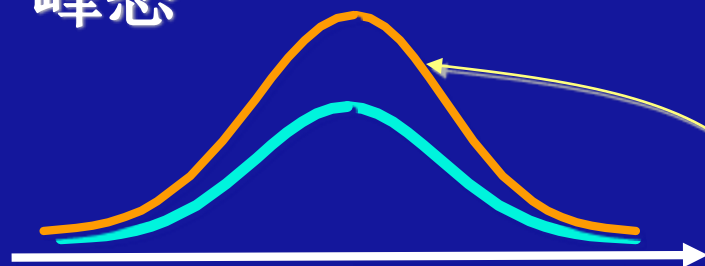


左偏分布

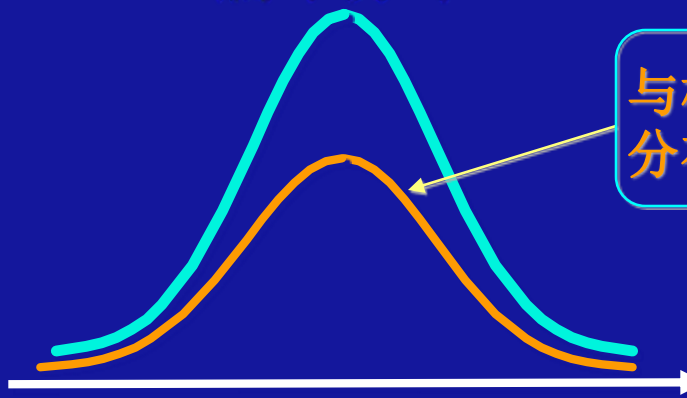


右偏分布

峰态



扁平分布

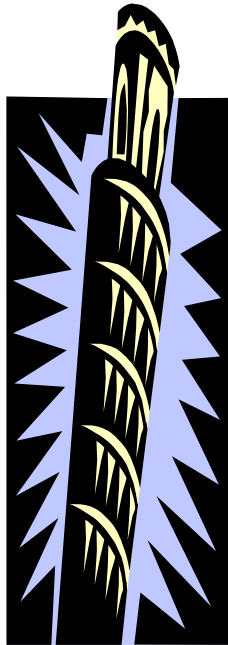


尖峰分布

与标准正态分布比较!

偏态 (skewness)

1. 统计学家**K.Pearson**于**1895**年首次提出。
是指数据分布的不对称性
2. 测度统计量是偏态系数 (**coefficient of skewness**)
3. 偏态系数**=0**为对称分布；**>0**为右偏分布；**<0**为左偏分布
4. 偏态系数大于**1**或小于**-1**，为高度偏态分布；偏态系数在**0.5~1**或**-1~-0.5**之间，为中等偏态分布；偏态系数越接近**0**，偏斜程度就越低



峰态 (kurtosis)

1. 统计学家**K.Pearson**于**1905**年首次提出。数据分布峰值的高低
2. 测度统计量是峰态系数(**coefficient of kurtosis**)
3. 峰态系数**=0**扁平峰度适中
4. 峰态系数**<0**为扁平分布
5. 峰态系数**>0**为尖峰分布



原始数据: 17 16 21 18 13 16 12 11

EXCEL输出结果:

平均	15.5
标准误差	1.180193689
中位数	16
众数	16
标准差	3.338091842
方差	11.14285714
峰度	-0.596449704
区域	10
最小值	11
最大值	21
求和	124
观测数	8



SPSS描述性分析

Statistics

N	Valid	8
	Missing	0
Mean		15.5000
Std. Error of Mean		1.18019
Median		16.0000
Mode		16.00
Std. Deviation		3.33809
Variance		11.143
Skewness		.184
Kurtosis		-.596
Range		10.00
Minimum		11.00
Maximum		21.00
Sum		124.00



位置测度指标

- 位次指标： 根据观察值在变量数列中的位置而确定的指标

- 常用的位置测度指标有：

四分位次指标

十分位次指标

百分位次指标

四分位距



五大位次指标

(三个四分位次指标加上最大值及最小值)

▼五大位次具体指：

◆最小值

◆第一四分位数

◆第二四分位数（中位数）

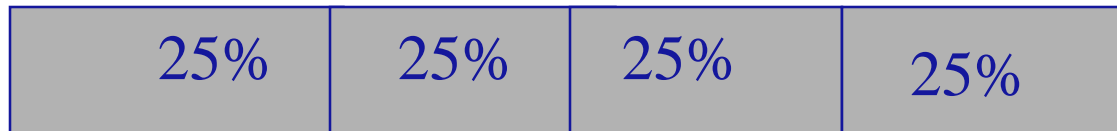
◆第三四分位数

◆最大值



五大位次指标位置的图示

- 把变量数列(从小到大排列)分成四等份



最小值 Q₁ 中位数 Q₃ 最大值



四分位数的确定1

原始数据: 10.3 4.9 8.9 11.7 6.3 7.7

按顺序排列: 4.9 6.3 7.7 8.9 10.3 11.7

位置: 1 2 3 4 5 6

第1四分位数的位置公式为:

$$Q1 \text{位置} = 1(n+1)/4$$



四分位数的确定2

❖ $Q_1 \text{ 位置} = 1(n+1)/4 = 1(6+1)/4 = 1.75$

$Q_1 = 6.3$

❖ $Q_2 \text{ 位置} = 2(n+1)/4 = 2(6+1)/4 = 3.5$

$Q_2 = (7.7 + 8.9)/2 = 8.3$

❖ $Q_3 \text{ 位置} = 3(n+1)/4 = 3(6+1)/4 = 5.25$

$Q_3 = 10.3$

四分位距

$Q_r = Q_3 - Q_1$ 用于说明中间**50%**数据的离散程度



四分位数

- ❖ 把一组数据按从小到大（或从大到小）的次序排成一个数列，将这个数列分成**4**个部分，每个部分包含数目相等的数据，各部分数据分界点上的数据值叫做四分位数。
- ❖ 第一个四分位数**Q1**之前包括了**25%**的数据，第二个四分位数**Q2**即中位数，中位数之前包括了**50%**的数据，第三个四分位数**Q3**之前包含了**75%**的数据。



四分位差

- ❖ 舍去数列中数值最高的**25%**数据和数值最低的**25%**的数据，求出中间**50%**的数据中最大数据与最小数据的数值差，即四分位差。
- ❖ 四分位差表明有**50%**的样本值在分布在这一区间内，用**Q** 代表四分位差，计算公式为：

❖
$$Q_r = Q_3 - Q_1$$



五大位次指标的图示：箱索图

❖箱索图是一种将五个位次指标显示在一条横轴上，以刻画变量数列集中、离散和偏斜态势的统计图

❖主要作用：

- 用于对两个或两个以上数列的集中、离散和偏斜态势作迅速而直观的对比。
- 识别数据中是否存在异常值



箱索图的画法

❖ 首先确定

- 第一四分位数、
- 第二四分位数（中位数）
- 第三四分位数

❖ 计算1.5倍的四分位距

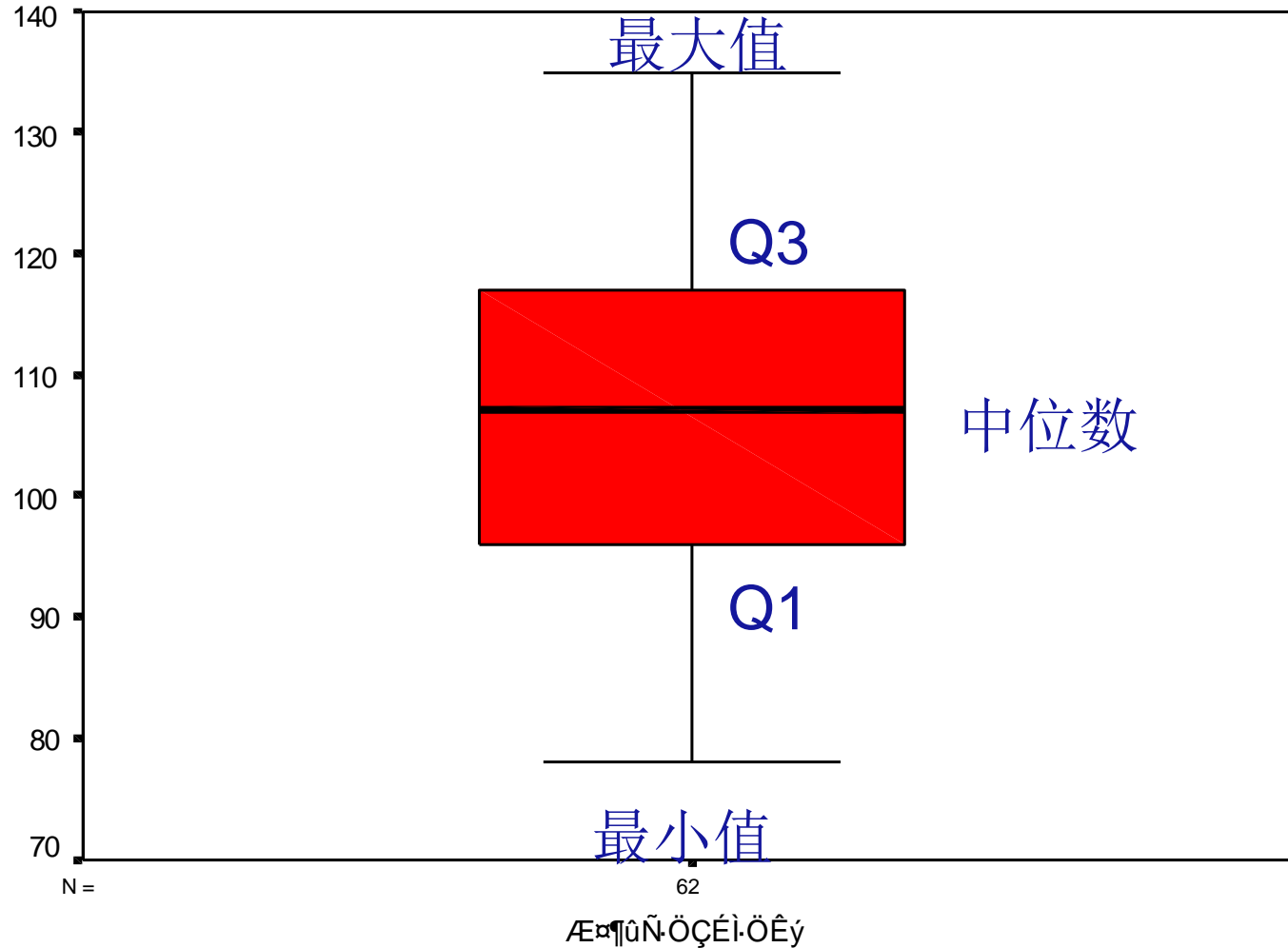


62人智商分数的箱索图

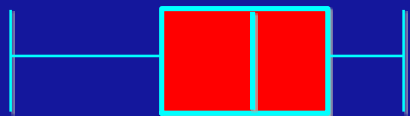
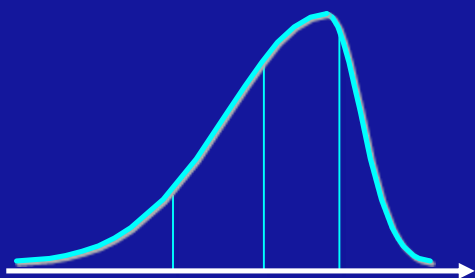
- 第一四分位数 **95.75**
- 第二四分位数（中位数） **107**
- 第三四分位数 **117.25**
- 四分位距 **21.5**
- $Q3 + 1.5 * 21.5 = 107.25 + 32.25 = 139.5$
- 最小值 = **78**
- 最大值 = **135**
- $Q1 - 1.5 * 21.5 = 95.75 - 32.25 = 63.25$



箱索图 -----探索数据分布规律的常用图形

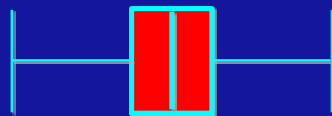
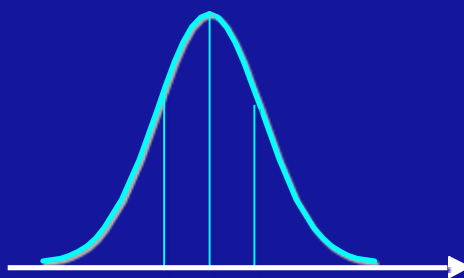


分布的形状与箱线图



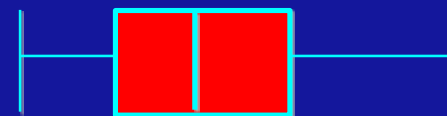
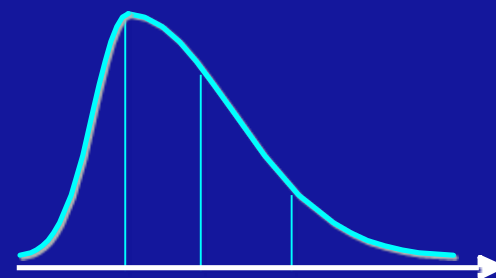
左偏分布

Left-skewed distribution



对称分布

Bell-shaped distribution



右偏分布

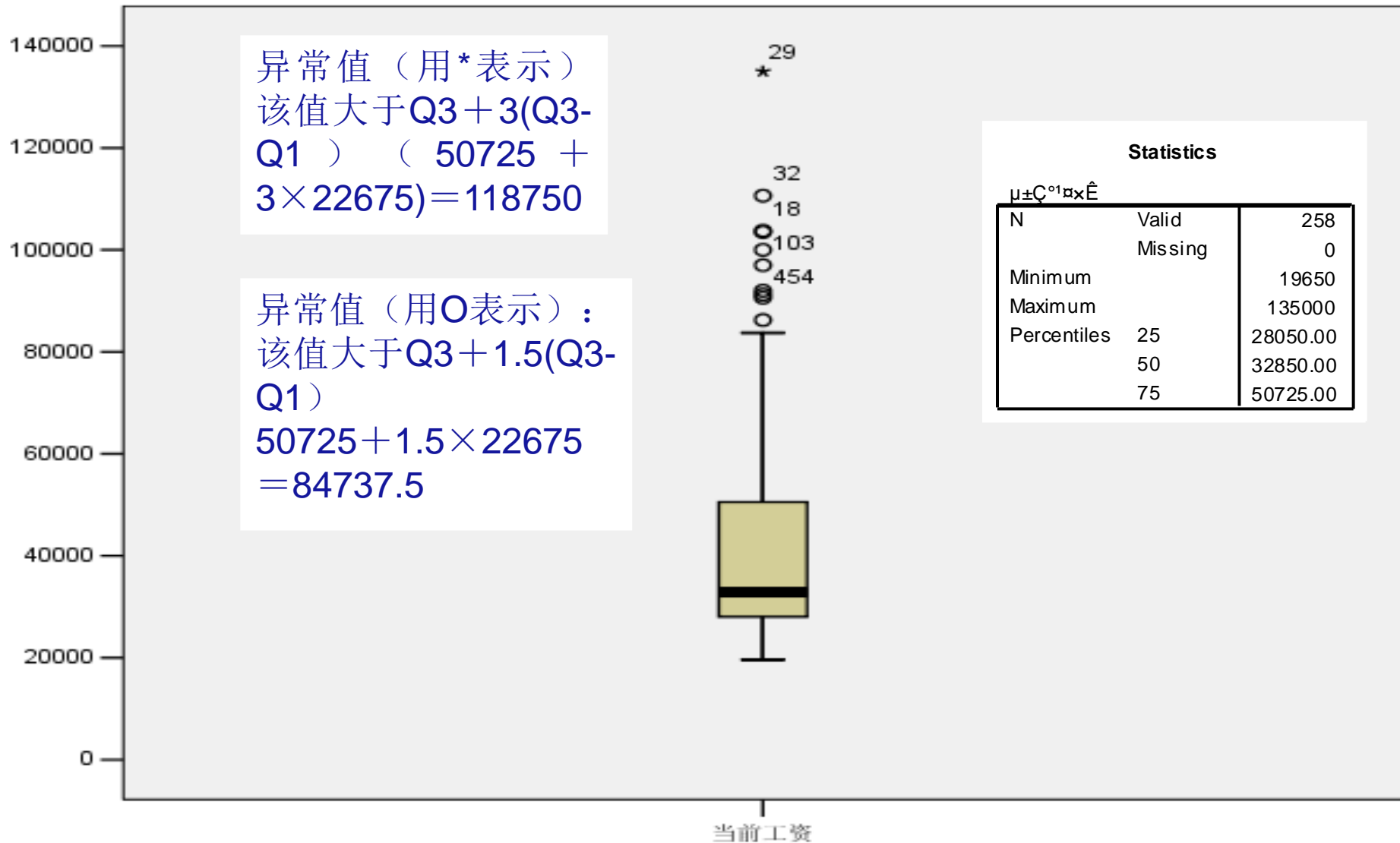
Right-skewed distribution

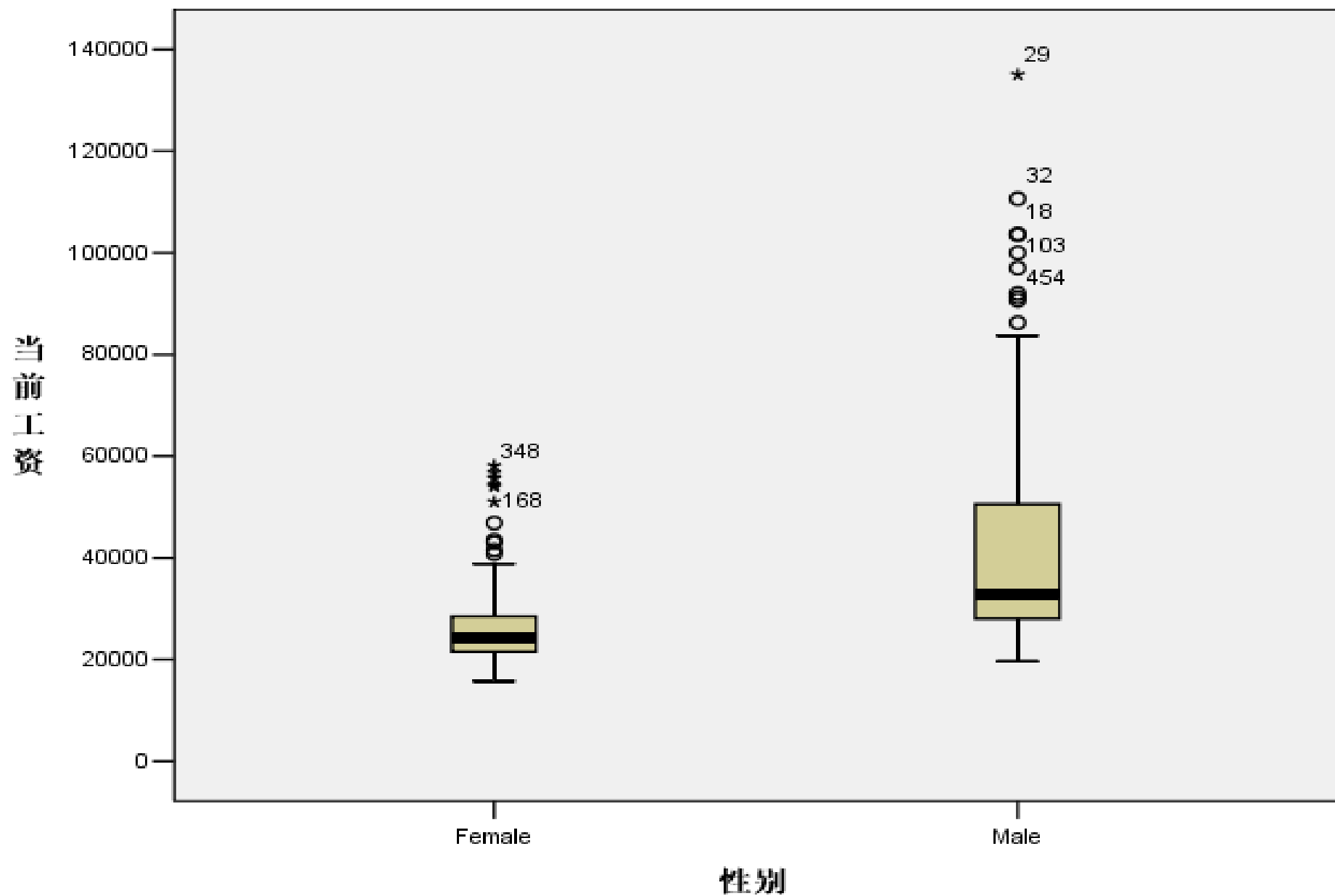
不同分布的箱线图

某厂男性职工年薪的箱索图

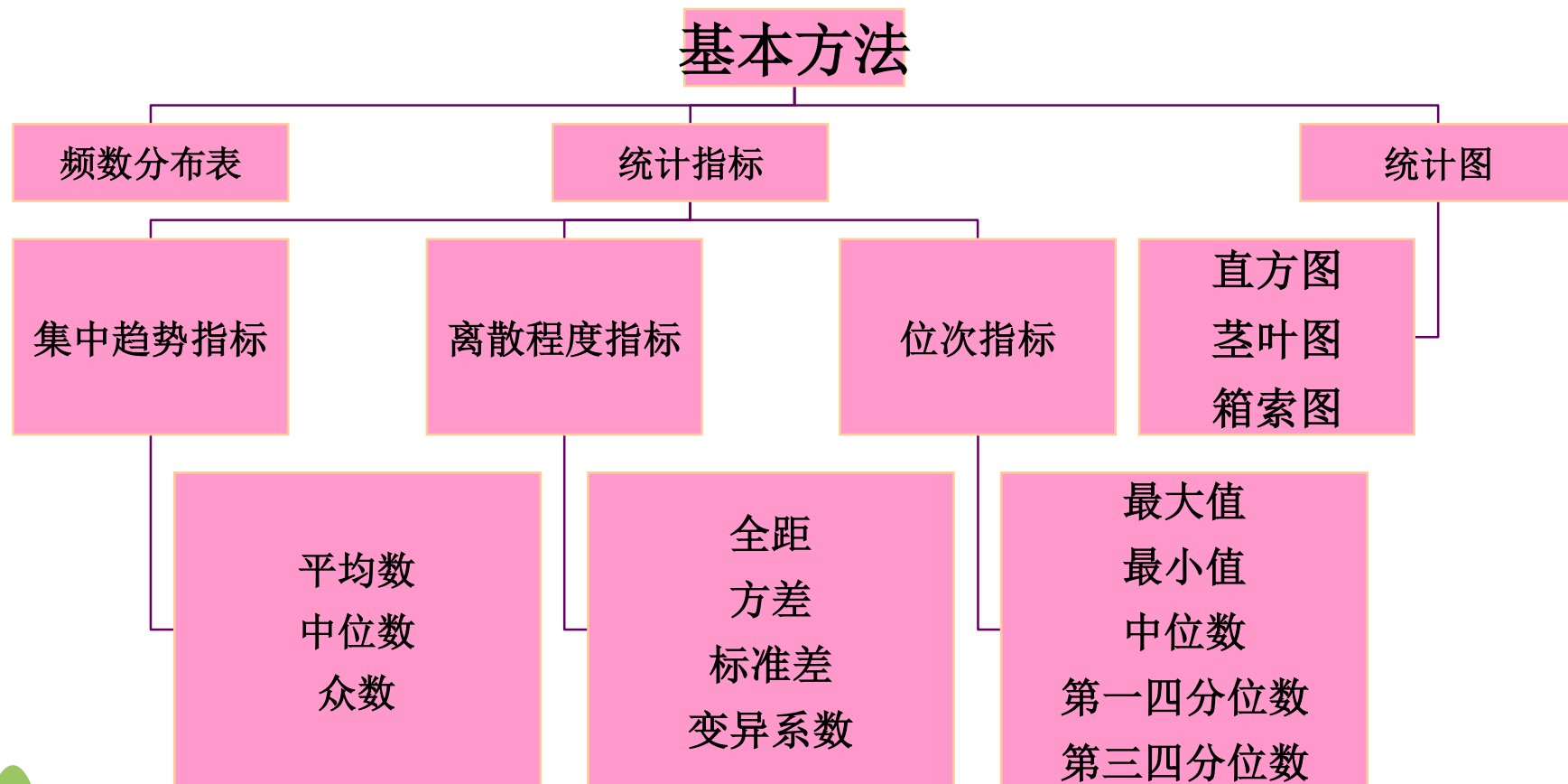
异常值（用*表示）
该值大于 $Q3 + 3(Q3 - Q1)$ （ $50725 + 3 \times 22675 = 118750$ ）

异常值（用O表示）：
该值大于 $Q3 + 1.5(Q3 - Q1)$
 $50725 + 1.5 \times 22675 = 84737.5$





探索、描述、分析单变量截面数据的基本统计方法



本讲要点回顾

- ❖ 熟练掌握描述数据分布中心的指标
 - 平均数、中位数、众数
- ❖ 熟练掌握描述数据分散程度的指标
 - 全距、方差、标准差、变异系数
- ❖ 熟练掌握描述 数据分布的位次指标
- ❖ 熟练掌握探索数据分布规律的常用图形



案例分析

“我看中日经济关系”数据库

期货投资数据的描述性统计分析



结束



谢谢!

[返回](#)

