2 描述统计

以表格、图形或数值形式描述统计
数据的各种特征，如同给人画像一样。

2.1 表格法
2.2 图形法
2.3 数值法
2.4 谨防统计“陷阱”

2.1 表格法（Tabular Methods）

例2.1.1：某家超市50次购买的软饮料名称

表2.1.1 50次购买的软饮料名称

Coke Classic	Coke Classic	Coke Classic	Dr. Pepper
Diet Coke	Diet Coke	Coke Classic	Coke Classic
Pepsi-Cola	Coke Classic	Pepsi-Cola	Diet Coke
Diet Coke	Coke Classic	Coke Classic	Pepsi-Cola
Coke Classic	Sprite	Sprite	Pepsi-Cola
Coke Classic	Coke Classic	Dr. Pepper	Pepsi-Cola
Dr. Pepper	Diet Coke	Pepsi-Cola	Pepsi-Cola
Diet Coke	Coke Classic	Diet Coke	Coke Classic
Pepsi-Cola	Diet Coke	Pepsi-Cola	Dr. Pepper
Pepsi-Cola	Coke Classic	Coke Classic	Pepsi-Cola
Coke Classic	Sprite	Coke Classic	Sprite
Dr. Pepper	Pepsi-Cola	Coke Classic	
Sprite	Coke Classic	Pepsi-Cola	

2.1 表格法（续）

例2.1.2：某公司20次年末审计时间

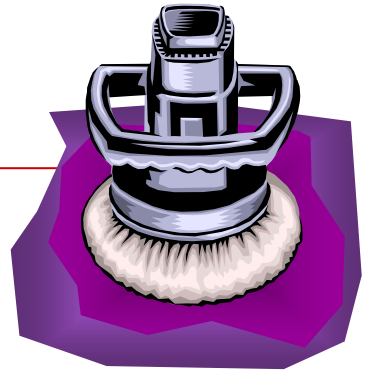
表2.1.2 20次年末审计时间（单位：小时）

12	14	21	22	17
15	14	18	33	23
20	15	19	16	28
22	27	18	18	13

问题2.1.1

- 这些数据的分布状况如何，该通过什么方式来描述？
- 这些数据的一般水平如何，该使用什么指标来描述？
- 这些数据的离散状况如何，该使用什么指标来描述？

2.1 表格法（续）



频数分布 (Frequency distribution)

基本概念

- **频数分布**：在统计分组的基础上，将总体（样本）所有单位按某一标志进行归纳排列，从而形成各个单位在各组间的分布。
- **双重含义**：对总体（样本）是“分”，对单位是“合”。
- **分组原则**：互斥、穷举。
- **由两个要素组成**：一是总体（样本）按某一标志所分的组；二是各组的单位数（频数、次数）。
- **定性数据频数分布的基本步骤**：确定组、计数。
- **定量数据频数分布的基本步骤**：确定组数、组距和组限、计数。

2.1 表格法（续）

➢ **软饮料的频数/相对/百分比/累积频数分布**

表2.1.3 软饮料的频数/相对/百分比/累积频数分布

软饮料	频数	相对频数	百分比频数	向上累积频数	向下累积频数
Coke lassic	19	0.38	38	19	50
Diet Coke	8	0.16	16	27	31
Dr.Pepper	5	0.10	10	32	23
Pesi-Cola	13	0.26	26	45	18
Sprite	5	0.10	10	50	5
合计	50	1.00	100	--	--

- **频数**：每组包含观测值的数目。
- **相对频数**：组频数/观测值总数；
- **百分比频数**：(组频数/观测值总数)×100%；
- **向上累积频数**：各组组上限以下的观测值数目。

2.1 表格法（续）

➢ 审计时间的频数/相对/百分比/累积频数分布

表2.1.4 审计时间的频数/相对/百分比/累积频数分布

审计时间	频数	相对频数	百分比频数	向上累积频数	向下累积频数
10—15	4	0.20	20	4	20
15—20	8	0.40	40	12	16
20—25	5	0.25	25	17	8
25—30	2	0.10	10	19	3
30—35	1	0.05	5	20	1
合计	20	1	100	--	--

- 遵循“上限不在组内”原则。
- 组距=组上限—组下限。
- 组中值=(组上限+组下限)/2。

0. 400 个工人每周工作的小时数如表 3.5 所示：

表 3.5 工人每周工作小时频数分布

小时	频数
0 - 10	20
10 -20	80
20 - 30	200
30 - 40	100

那么，不大于 29 小时的累积百分频数为_____，大于 29 小时的累积百分频数为_____。

2.1 表格法（续）

➢ 饭店的质量等级与餐价的交叉分组表

表2.1.5 饭店的质量等级与餐价的交叉分组表

	Meal Price				
Quality Rating	20以下	20～30	30～40	40以上	Total
Good	42	40	2	0	84
Very Good	34	64	46	6	150
Excellent	2	14	28	22	66
Total	78	118	76	28	300

2.1 表格法（续）

统计表的组成

总标题

表2.1.6　1997～1998年城镇居民家庭抽样调查资料

项目	单位	1997年	1998年
一、调查户数	户	37890	39080
二、平均每户家庭人口数	人	3.19	3.16
三、平均每户就业人口数	人	1.83	1.80
四、平均每人全部收入	元	5188.54	5458.34

列标题

行标题

数字资料

主词栏

宾词栏

资料来源：《中国统计摘要1999》，中国统计出版社，第79页。

注：1．本表为城市和县城的城镇居民家庭抽样调查材料。

2．消费性支出项目包括：食品、衣着、家庭设备用品及服务、医疗保健、交通和通讯、娱乐教育文化服务、居住、杂项商品和服务。

附注

2.2 图形法（Graphical Methods）

柱状图(bar graph)

定义

- 又称条形图，用相隔相同距离、具有固定宽度的条形来表示频数（或相对频数、百分比频数）分布的一种统计图。

绘制步骤

- 第1步，把原始数据进行分组，编制频数分布；
- 第2步，选取组的标志为横轴，组频数（相对频数、百分比频数）为纵轴，合适选择坐标刻度，并将每组的标志值及频数（相对频数、百分比频数）分别标记在横轴和纵轴上；
- 第3步，在每组标志值的标记上绘制条形，条形高度等于该组对应的频数（相对/百分比频数），而且保持每个条形宽度一样，间距也一样。

2.2 图形法（续）

➢ 软饮料的柱状图

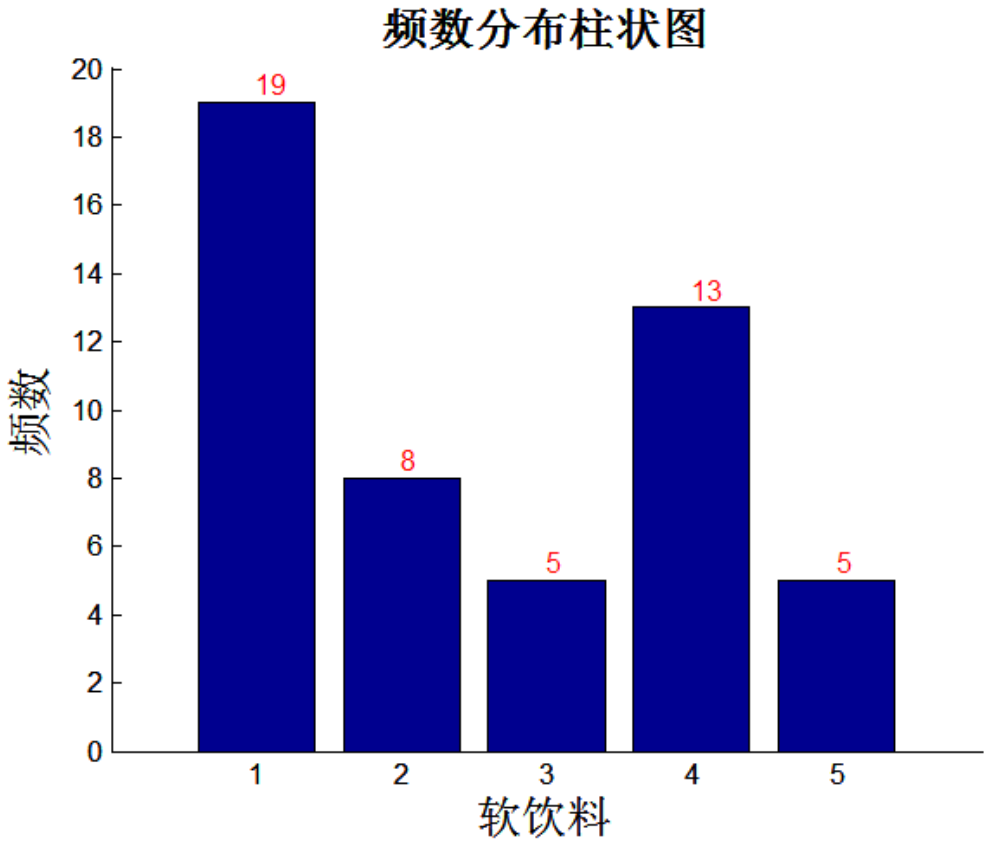


图2.2.1 软饮料的柱状图

2.2 图形法（续）

直方图(histogram)

定义

- 用一组没有间隔的矩形来表示表示频数（或相对频数、百分比频数）的一种统计图。

绘制步骤

- 第1步，对数据进行分组，编制频数分布；
- 第2步，选取组的标志为横轴，组频数（相对频数、百分比频数）为纵轴，合适选择坐标刻度，并将每组的标志值及频数（相对频数、百分比频数）分别标记在横轴和纵轴上；
- 第3步，绘制矩形，每个矩形代表一个组，矩形的底部宽度对应该组的组距，矩形的高度对应该组的频数（或相对频数、百分比频数），矩形之间不要有间距，除非该组是空的。

2.2 图形法（续）

➢ 审计时间的直方图

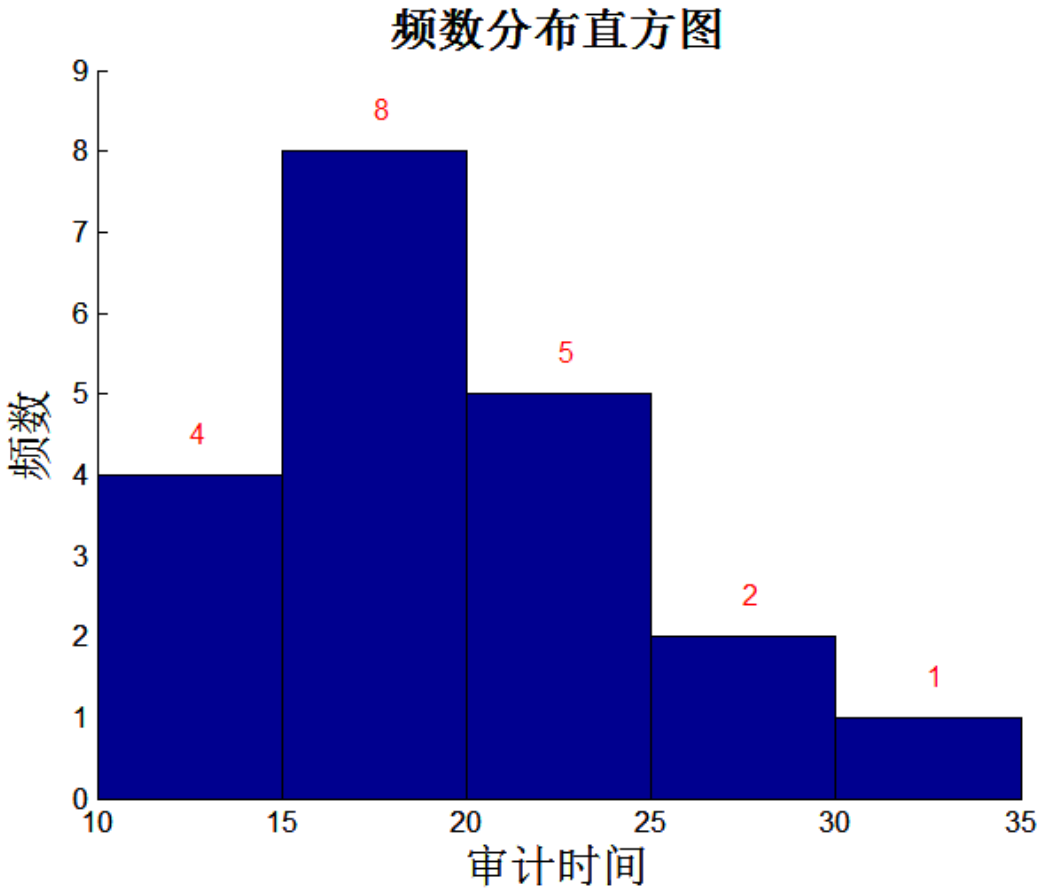


图2.2.2 审计时间的直方图

2.2 图形法（续）

饼图(pie chart)

定义

- 用圆面的扇形面积比率表示相对频数（或百分比频数）的一种统计图。

绘制步骤

- 第1步，画一个圆面；
- 第2步，根据相对频数（或百分比频数）计算出对应的扇形所包含的角度；
- 第3步，绘制出相应的扇形区域并进行有关标注。

2.2 图形法（续）

- 软饮料的饼图

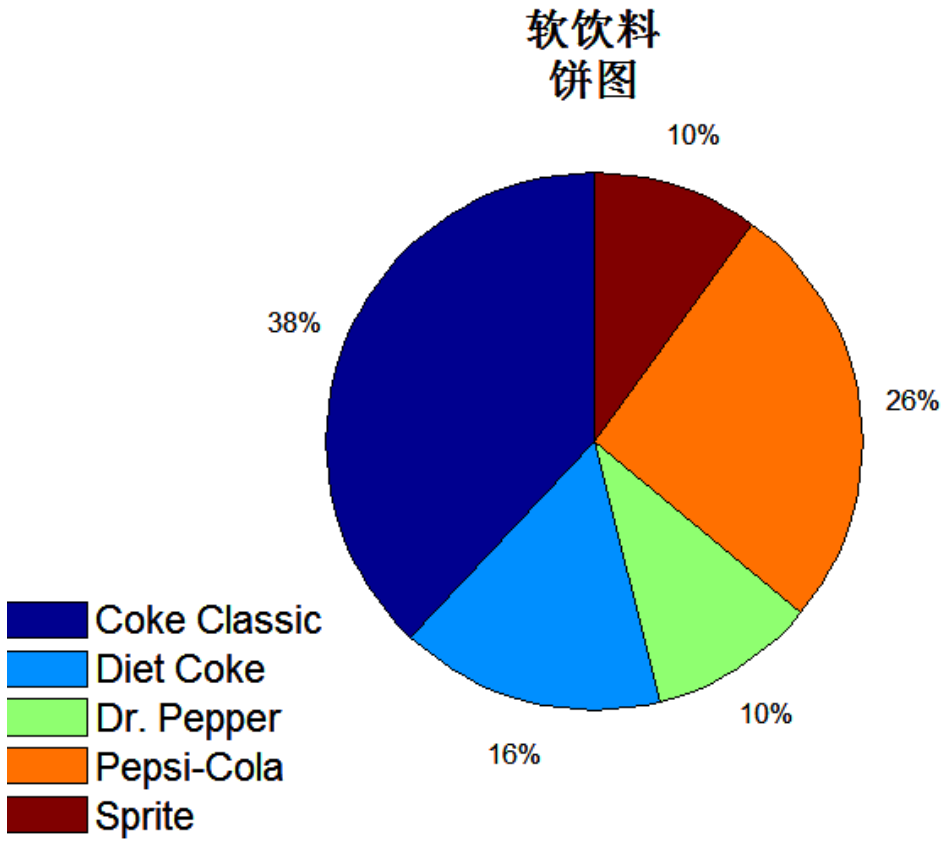


图2.2.3 软饮料的饼图

2.2 图形法（续）

- 审计时间的饼图

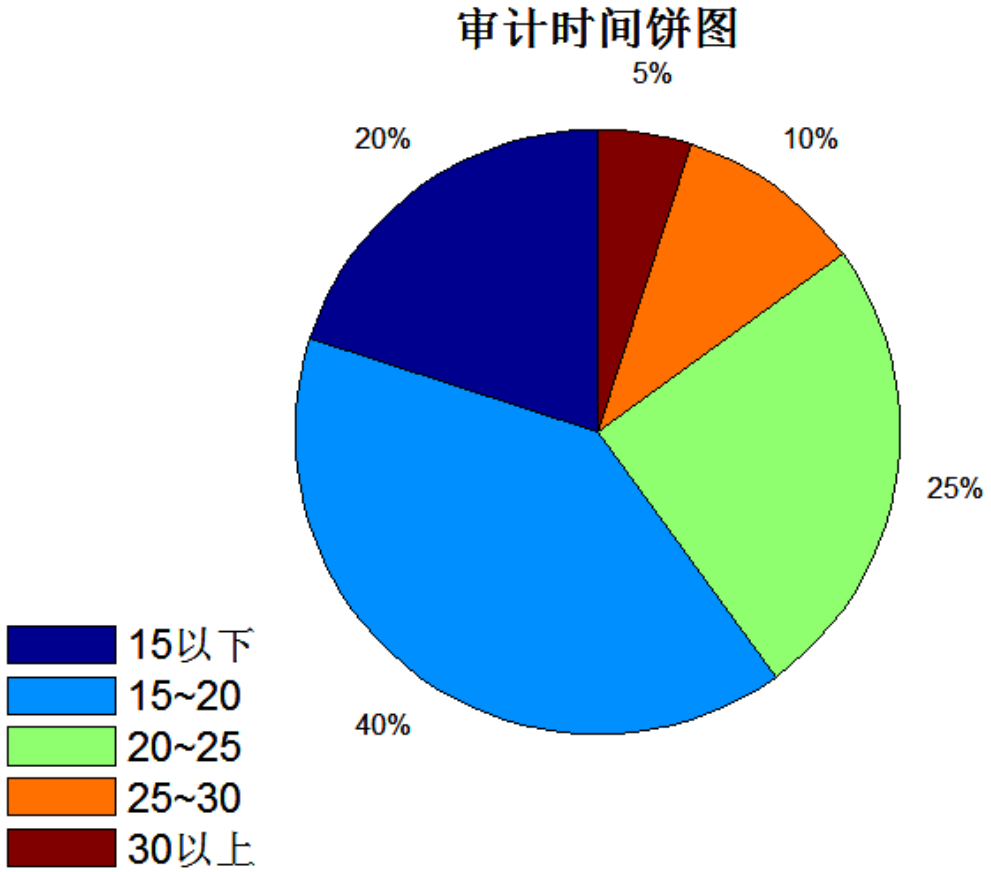
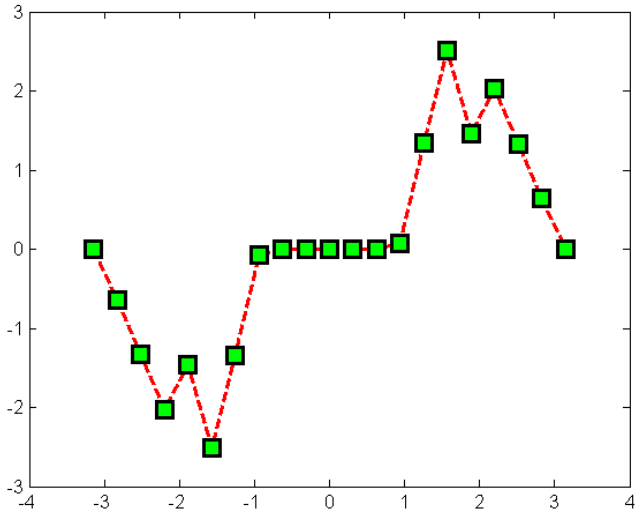
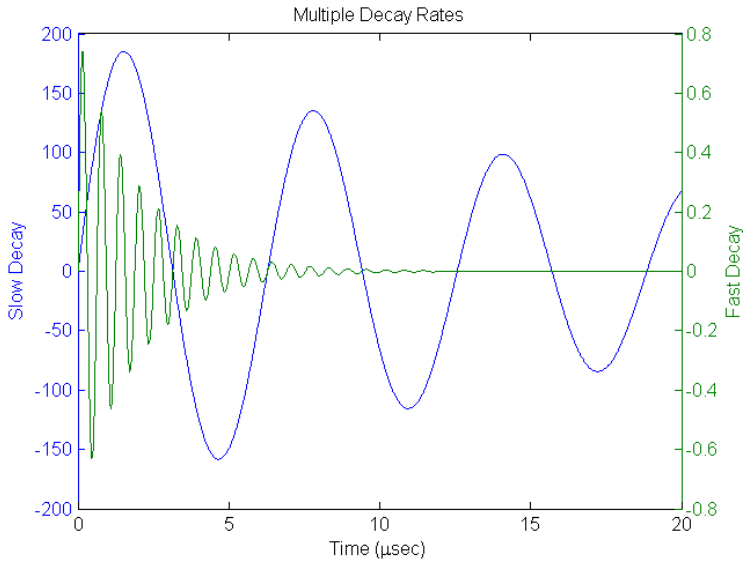
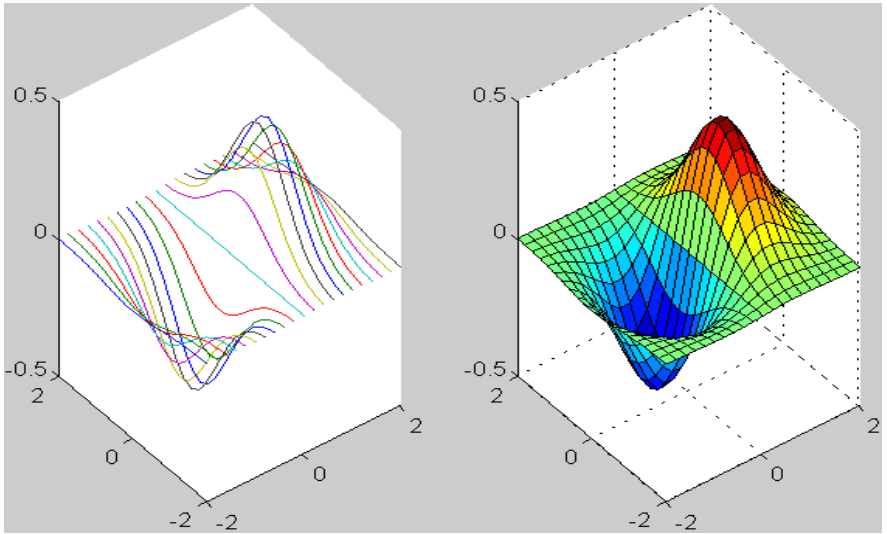


图2.2.4 审计时间的饼图

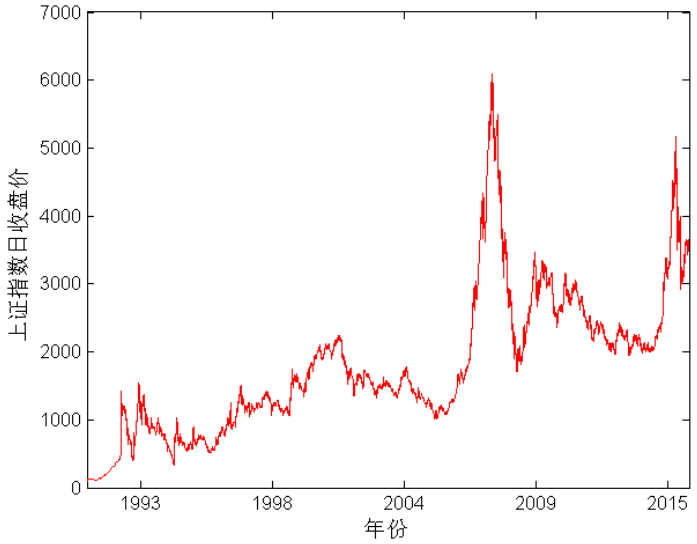
– 二维图与三维图

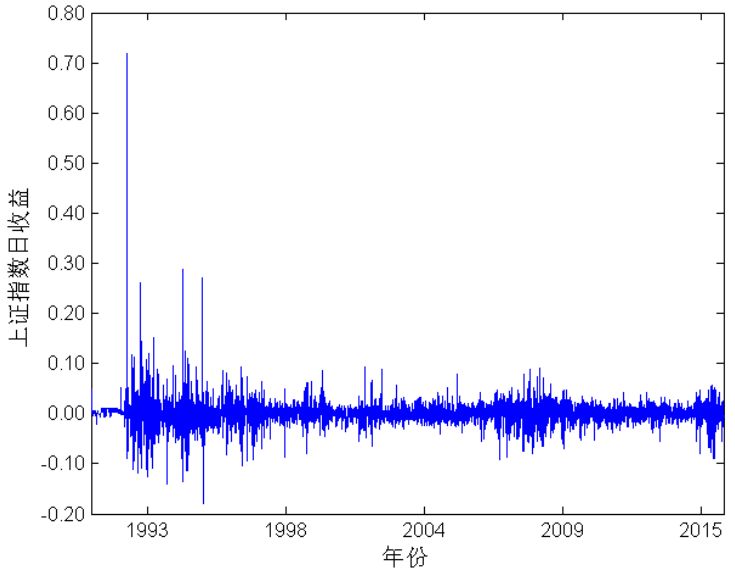

3
2
1
0
-1
-2
-3
-4
-3
-2
-1
0
1
2
3
4

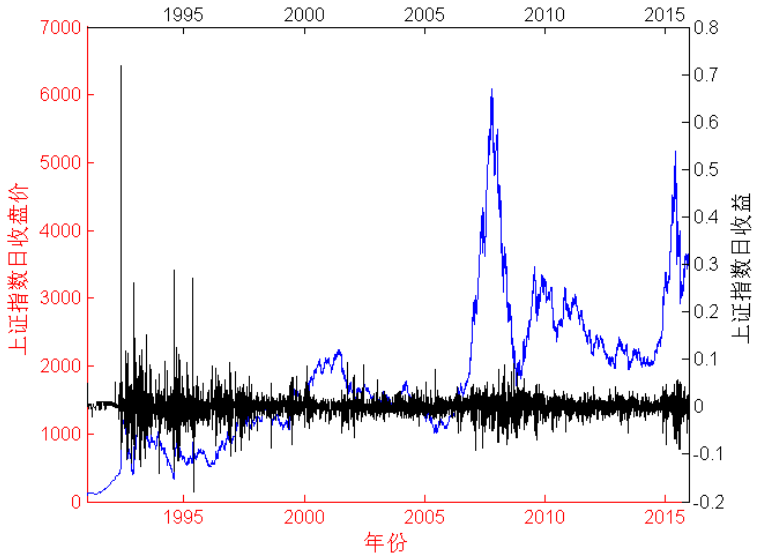

Multiple Decay Rates
Slow Decay
Fast Decay
Time (µsec)
200
150
100
50
0
-50
-100
-150
-200
0.8
0.6
0.4
0.2
0
-0.2
-0.4
-0.6
-0.8
0
5
10
15
20

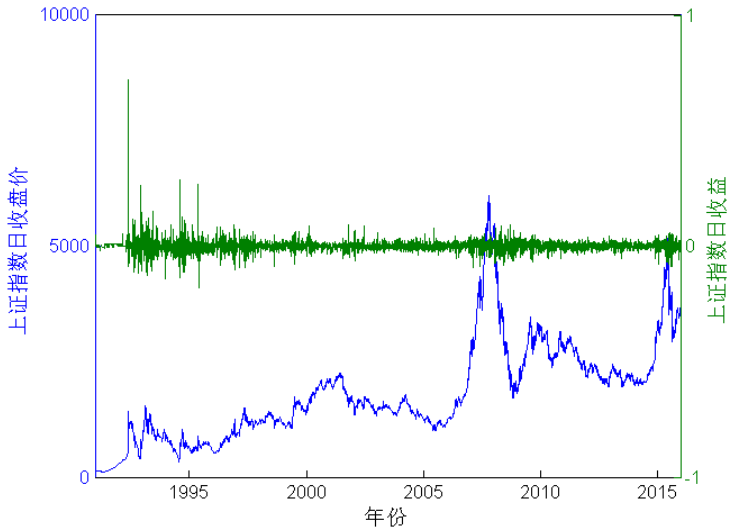

0.5
0
-0.5
2
0
-2
-2
0
2
0.5
0
-0.5
2
0
-2
-2
0
2

– 走势图

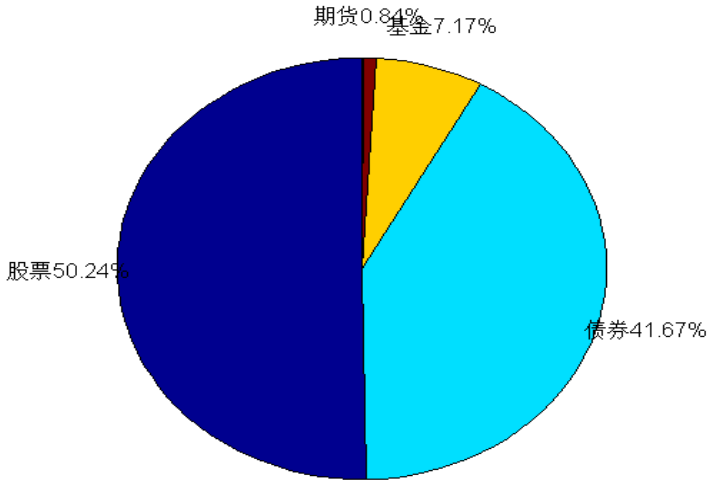

上证指数日收盘价
年份

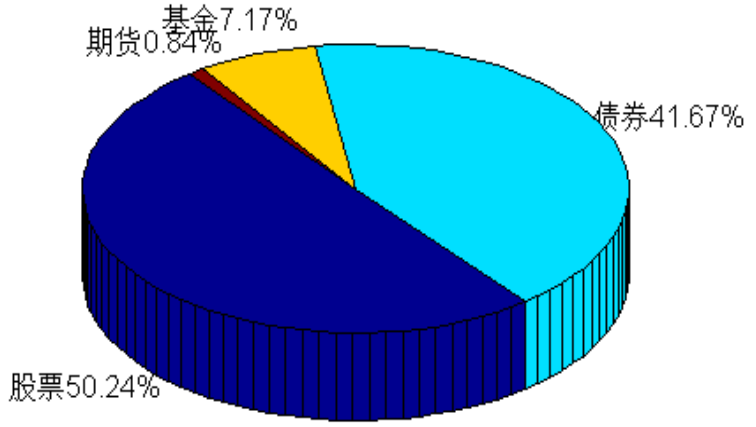

上证指数日收益
年份

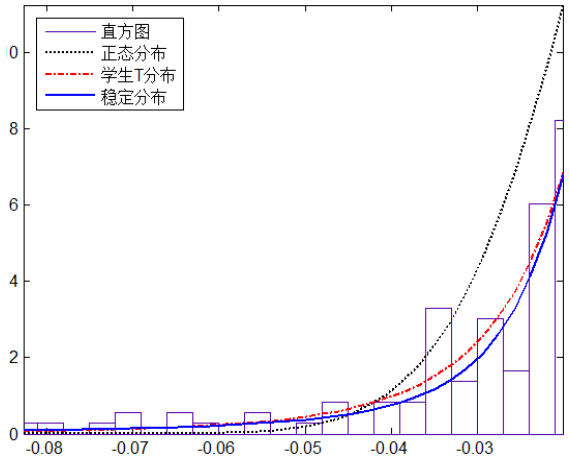

上证指数日收盘价
上证指数日收益
年份

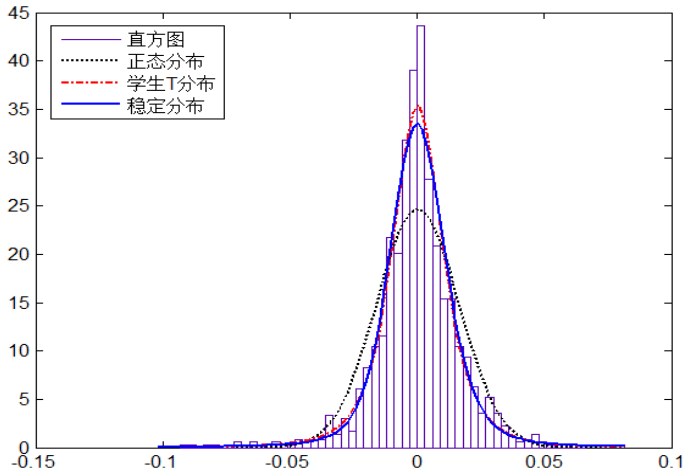

上证指数日收盘价
上证指数日收益
年份

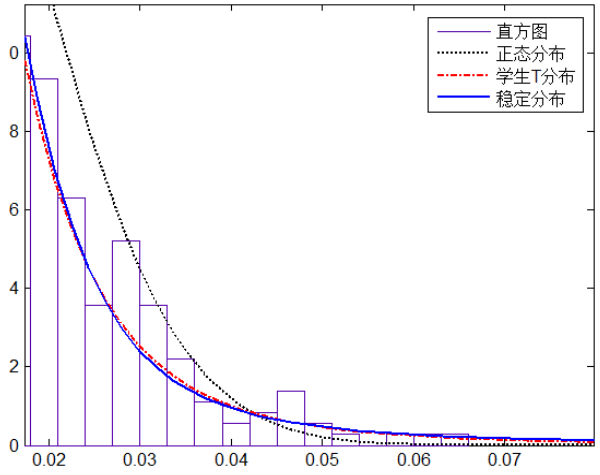
– 统计图


期货0.84%
基金7.17%
股票50.24%
债券41.67%

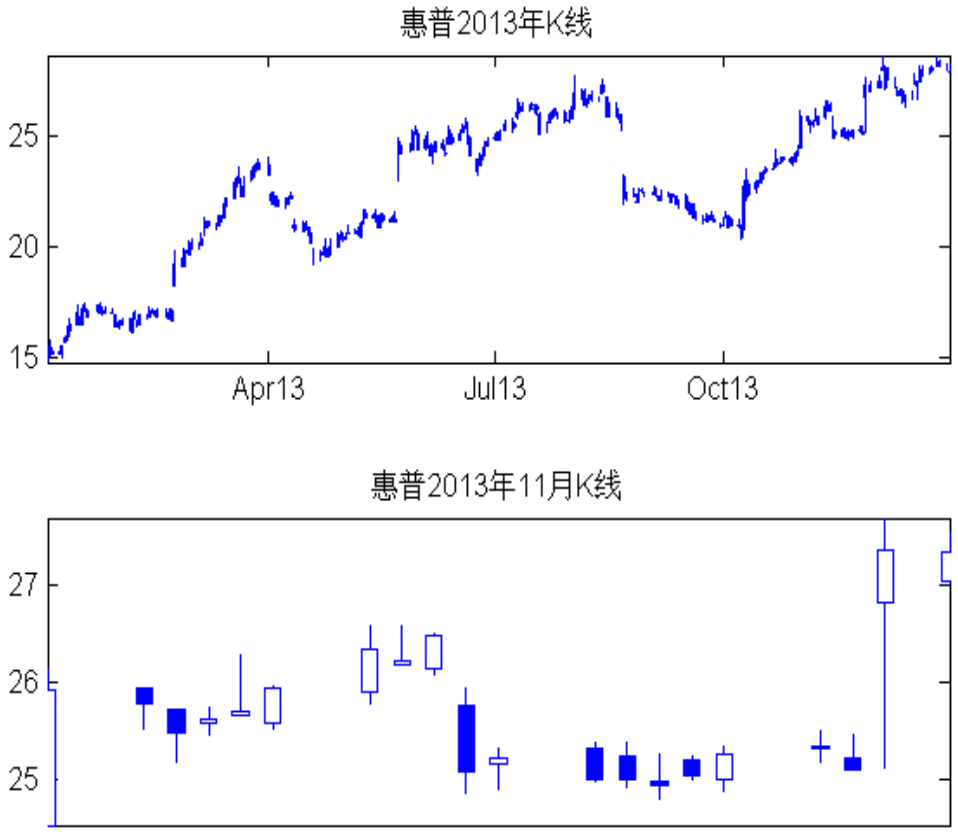

基金7.17%
期货0.84%
债券41.67%
股票50.24%

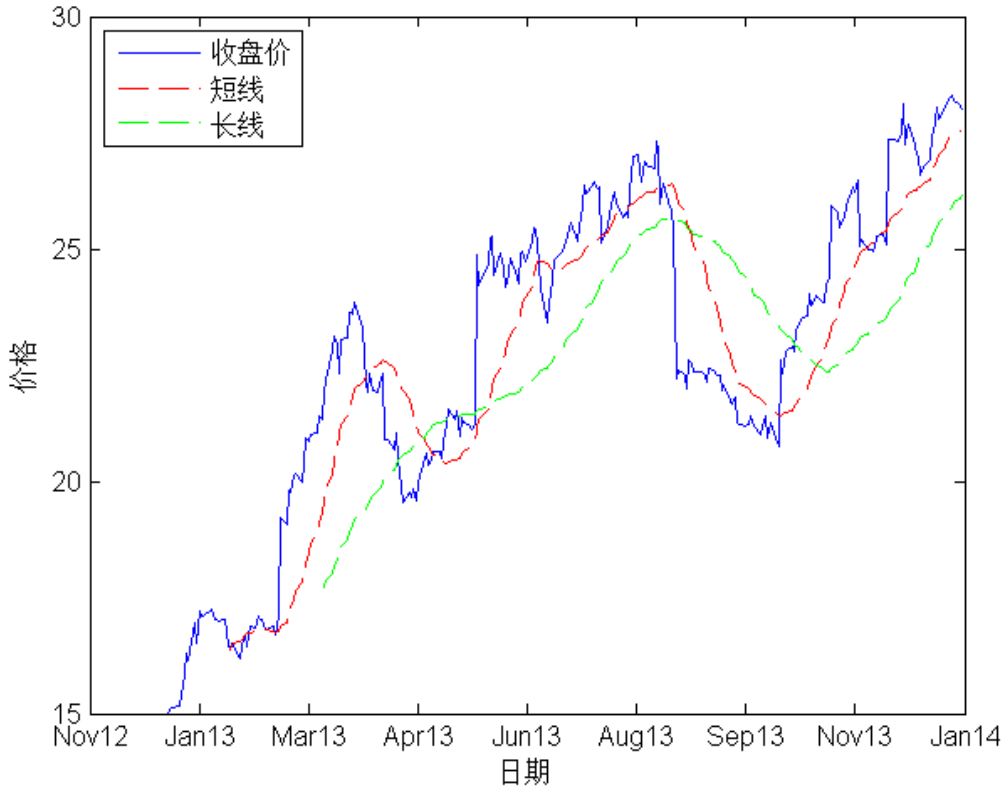

直方图
正态分布
学生T分布
稳定分布


直方图
正态分布
学生T分布
稳定分布

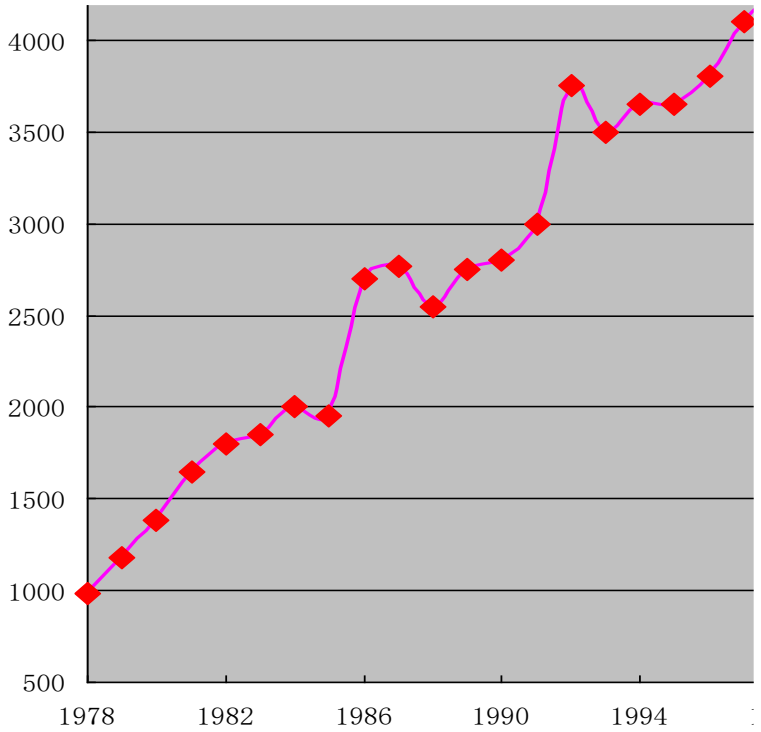

直方图
正态分布
学生T分布
稳定分布

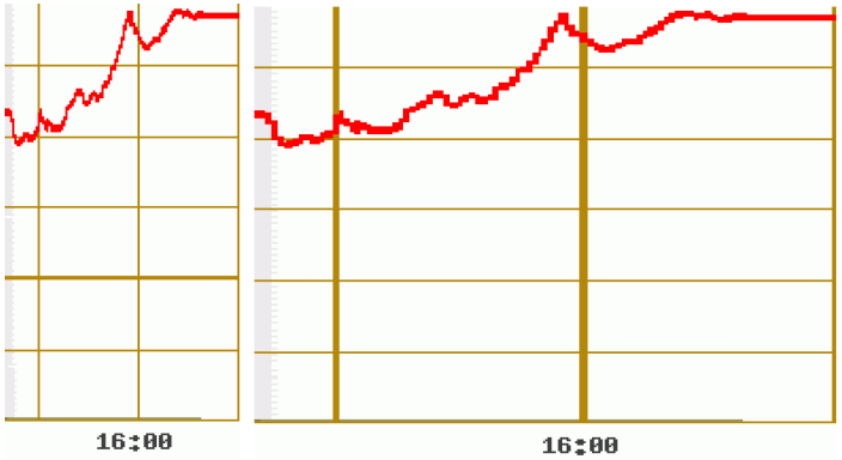
– 技术图

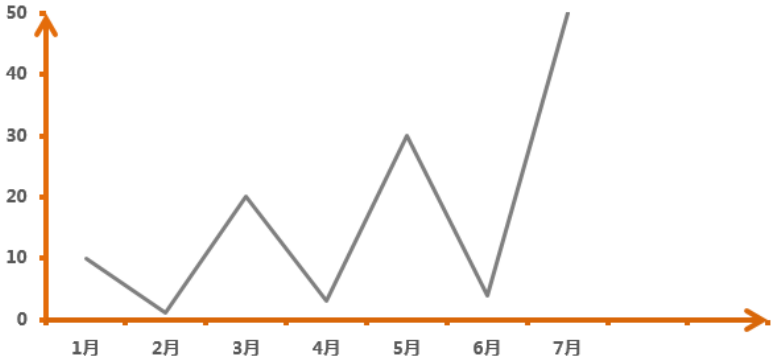

惠普2013年K线
25
20
15
Apr13
Jul13
Oct13
惠普2013年11月K线
27
26
25

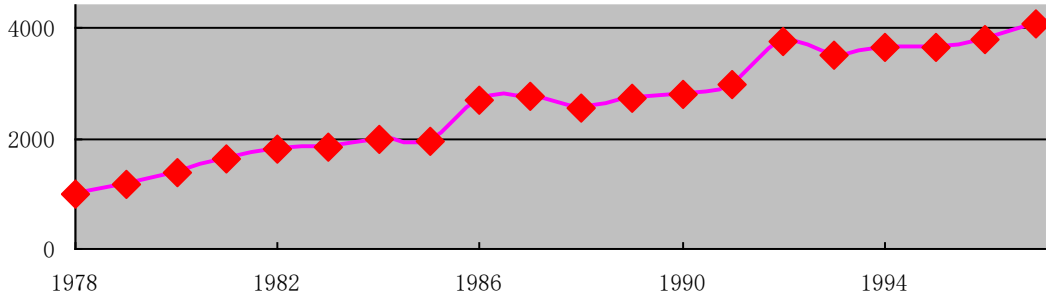

收盘价
短线
长线
30
25
20
15
价格
Nov12
Jan13
Mar13
Apr13
Jun13
Aug13
Sep13
Nov13
Jan14
日期

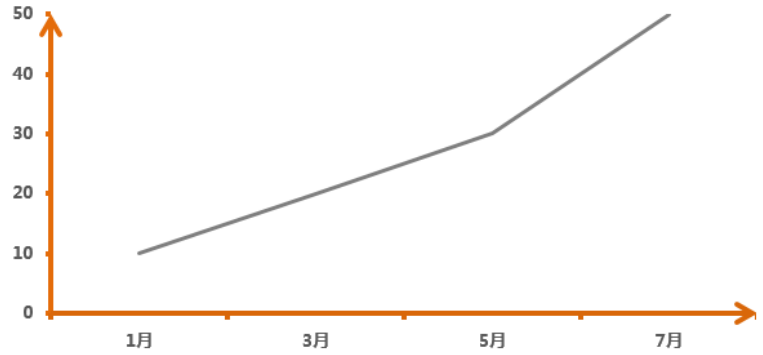
– 避免图形陷进


4000
3500
3000
2500
2000
1500
1000
500
1978
1982
1986
1990
1994


16:00
16:00


50
40
30
20
10
0
1月
2月
3月
4月
5月
6月
7月


4000
2000
0
1978
1982
1986
1990
1994


50
40
30
20
10
0
1月
3月
5月
7月

2.3 数值法（Numerical Methods）

概述

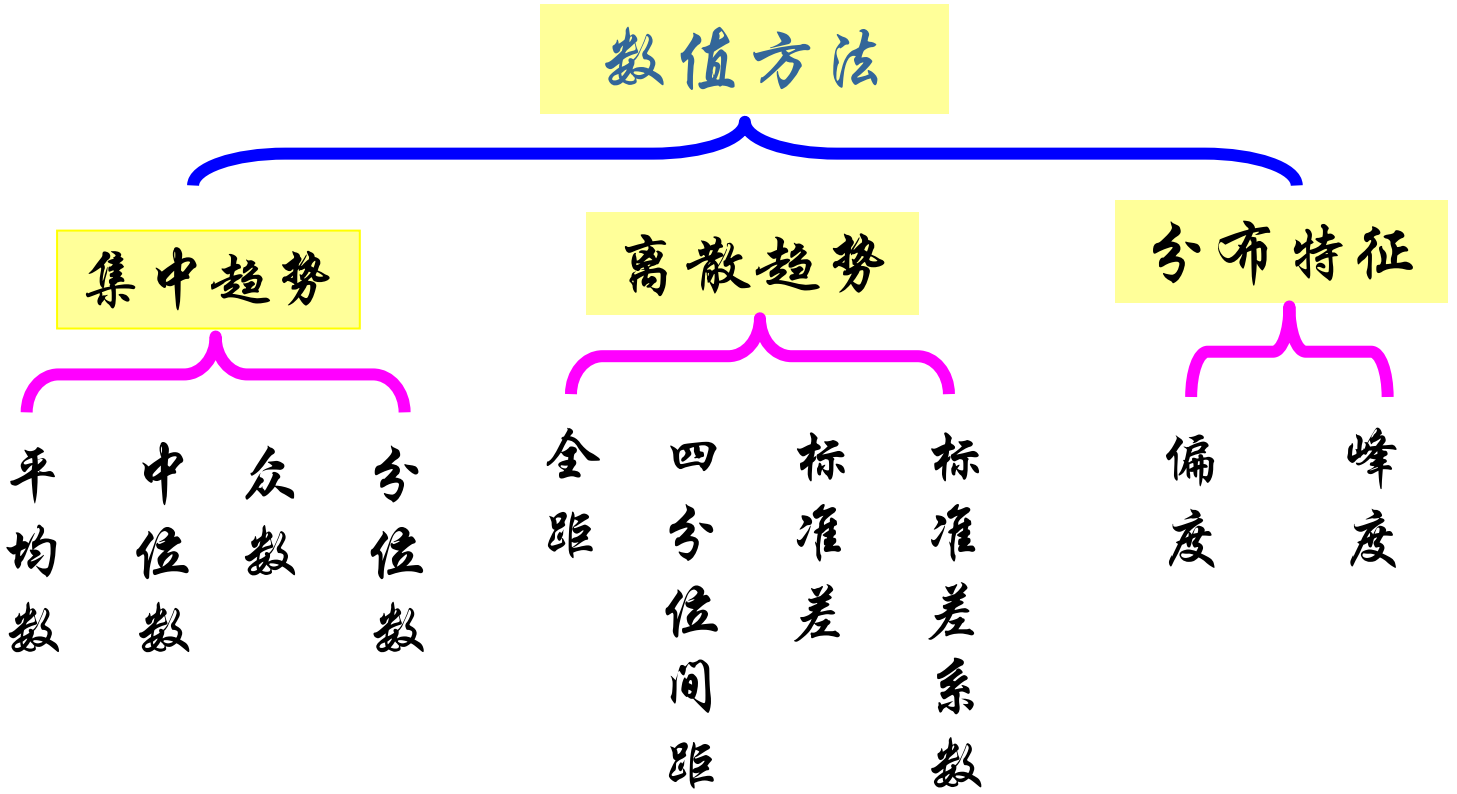


图2.3.1 描述统计的数值法

2.3 数值法（续）

算术平均数（Arithmetical Mean）

定义

- 标志总量与单位总数的比值，反映变量取值的一般水平和集中趋势，也通常称为平均值，简称均值，总体均值一般记为μ，样本均值一般记为$\bar{x}$。

公式

- **简单算术平均数：**

$$\mu = \frac{\sum_{i=1}^{N} x_i}{N} \quad ，或者 \quad \bar{x} = \frac{\sum_{i=1}^{n} x_i}{n}$$

其中，x_i为第i个单位的标志值，N为总体单位总数，n为样本容量。简单算术平均数适合未分组数据。

2.3 数值法（续）

- 加权算术平均数：

$$\mu = \frac{\sum_{i=1}^{M} M_i f_i}{\sum_{i=1}^{M} f_i} \quad ，或者 \quad \bar{x} = \frac{\sum_{i=1}^{m} M_i f_i}{\sum_{i=1}^{m} f_i}$$

其中，M_i为第i组的组中值，f_i 为第i组的频数（权数），M为总体组数，m为样本组数。加权算术平均数适合于分组数据。

➢ 特点

- 易受极端值影响；
- 当数据变化均匀时，适合度量数据集中趋势。

2.3 数值法（续）

➢ 审计时间的算术平均数

- **简单算术平均数：**未分组数据（表2.1.2）

$$\bar{x}=\frac{\sum_{i=1}^{n}x_i}{n}=\frac{12+14+21+\cdots+18+18+13}{20}$$
$$=19.25(\text{小时})$$

➢ 问题2.3.1

- 为什么未分组、单项式分组、组距式分组计算的算术平均数结果不同？

- 平均数（Mean）除算术平均数之外还有哪些形式的平均数？它们又表示什么实际意义？

2.3 数值法（续）

中位数（Me, median）

- 定义
  - 当数据按一定顺序（如由小到大）排列后，排在中间位置的数据值。分下列两种情况：

  如果 n 是奇数，即为序列中间位置(n+1)/2对应的数值；如果 n 是偶数，则为两个中间位置n/2和n/2+1对应数值的均值。

- 特点
  - 不易受到极端值影响；
  - 当含有异常值时，更适合度量数据的集中趋势。
- 问题2.3.2
  - 描述收入和财富的平均水平或集中趋势一般应选择哪个指标？（是平均数还是中位数）

2.3 数值法（续）

- 审计时间的中位数
  - **未分组数据：**

第1步，将原始数据从小到大排列：12, 13, 14, 14, 15, 15, 16, 17, 18, 18, 18, 19, 20, 21, 22, 22, 23, 27, 28, 33；

第2步，中间位置10和11对应的数值分别为18和18；

第3步，中位数为中间位置数值的平均数，即为18。

2.3 数值法（续）

众数（Mo, mode）

- 定义
  - 数据中出现次数最多的数据值，可能没有众数，也可能出现多众数现象。
- 审计时间的众数
  - **未分组数据：**出现次数最多为3次，众数为18。

2.3 数值法（续）

分位数

百分位数（Percentile）

- **定义**：第p百分位数是这样的数值：至少p%的观测值不比它大，即不大于它的观测值个数占全部观察测个数的比例为p%。
- **计算方法**：

第1步，按升序排列数据；

第2步，计算位置指数i=(p/100)×(n+1)，其中，n为观测值个数；（注意与中位数的确定方法是一致的）

第3步，采用线性插值法计算百分位数。

$$y = f(x) = y_1 + \frac{(x - x_1)}{(x_2 - x_1)}(y_2 - y_1)$$

2.3 数值法（续）

- **四分位数（Quartiles）：特殊的百分位数**
  - 第1四分位数Q_1，即为第25百分位数；
  - 第2四分位数Q_2，即为第50百分位数；
  - 第3四分位数Q_3，即为第75百分位数。
- **审计时间的百分位数和四分位数**

  12, 13, 14, 14, 15, 15, 16, 17, 18, 18, 18, 19, 20, 21, 22, 22, 23, 27, 28, 33
  - 10百分位数的位置指数为i=(10/100)×(20+1)=2.1，即为13+(14-13)×0.1=13.1；
  - 85百分位数的位置指数为i=(85/100)×(20+1)=17.85，即为第17与第18的平均：23+(27-23) × 0.85=26.4；
  - 第1四分位数Q_1的位置指数为i=(25/100)×(20+1)=5.25，即15+(15-15)× 0.25=15。

2.3 数值法（续）

➢ 四分位数的其他3种确定方法

- 第1种：按n/4、2n/4、3n/4分别确定第1、2、3四分位数的位置，即为前述方法；
- 第2种：按(n+1)/4、2(n+1)/4、3(n+1)/4分别确定第1、2、3四分位数的位置，诸如Minitab软件采用此方法；
- 第3种：按(n+3)/4、(2n+2)/4、(3n+1)/4分别确定第1、2、3四分位数的位置，诸如Excel软件采用此方法。

2.3 数值法（续）

全距 （Range）

- 定义
  - 所有数据中最大值与最小值之差，即R=Max (x_i)-Min (x_i)。
- 特点
  - 极易受到端值的影响。

四分位间距（IQR）

- 定义
  - 第3个四分位数与第1个四分位数之差，即IQR=Q_3－Q_1。
- 特点
  - 可避免异常值的影响。

2.3 数值法（续）

方差（Variance）/标准差（ Standard Deviation ）

定义

- **方差**：所有单位标志值与其平均数的离差平方的算术平均数，总体方差一般记为σ^2，样本方差一般记为s^2。
- **标准差**：方差的正平方根，即$\sigma=\sqrt{\sigma^2}$或$s=\sqrt{s^2}$
- 均反映所有标志值与其平均数之间的平均差异程度，也反映所有标志值的离散趋势。

公式

- **简单方差**：适合未分组数据

$$\sigma^2=\frac{\sum_{i=1}^{N}(x_i-\mu)^2}{N}\quad，或者\quad s^2=\frac{\sum_{i=1}^{n}(x_i-\bar{x})^2}{n-1}$$

2.3 数值法（续）

- 特点
  - 方差和标准差均易受标志值计量单位和平均数大小的影响；
  - 方差不利于与平均数和原始数据比较，而标准差与平均数和原始数据同度量单位，有利于比较。

2.3 数值法（续）

➢ 审计时间的方差和标准差

- 未分组数据：

$$s^2=\frac{\sum_{i=1}^{n}(x_i-\bar{x})^2}{n-1}=\frac{(12-19.25)^2+\cdots+(33-19.25)^2}{20-1}=29.57$$

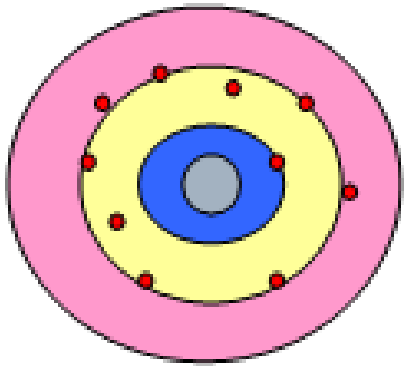
$$s=\sqrt{29.57}=5.44$$

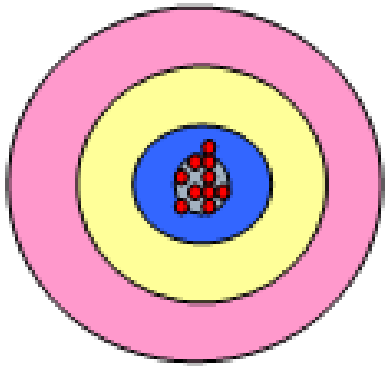
- 分组数据：表2.1.4

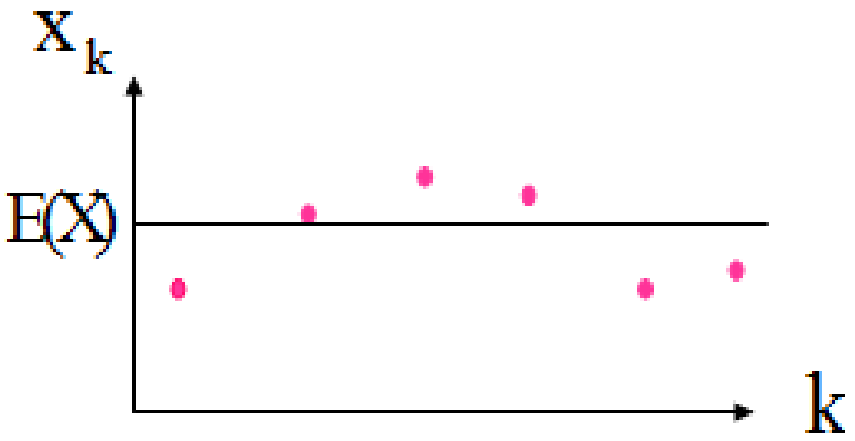
$$s^2=\frac{\sum_{i=1}^{m}(M_i-\bar{x})^2 f_i}{\sum_{i=1}^{m} f_i-1}=\frac{(12.5-19.5)^2+\cdots+(32.5-19.5)^2}{20-1}=30$$

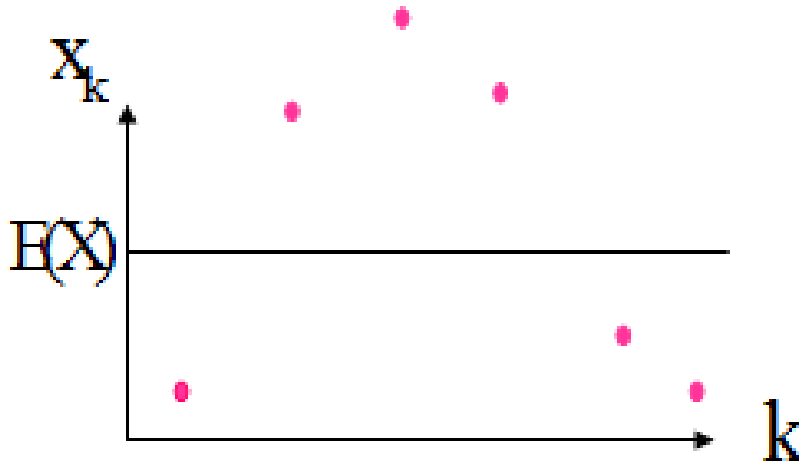
$$s=\sqrt{30}=5.48$$

哪个表示的标准差更大？









- 比较三分布的工程损失期望值和标准差。

三个损失分布

分布1		分布2		分布3	
损失结果	概率	损失结果	概率	损失结果	概率
250美元	0.33	0美元	0.33	0美元	0.4
500美元	0.34	500美元	0.34	500美元	0.2
750美元	0.33	1000美元	0.33	1000美元	0.4

2.3 数值法（续）

标准差系数（Coefficient of Standard Deviation ）

定义

- 标准差与平均数的比值。

公式

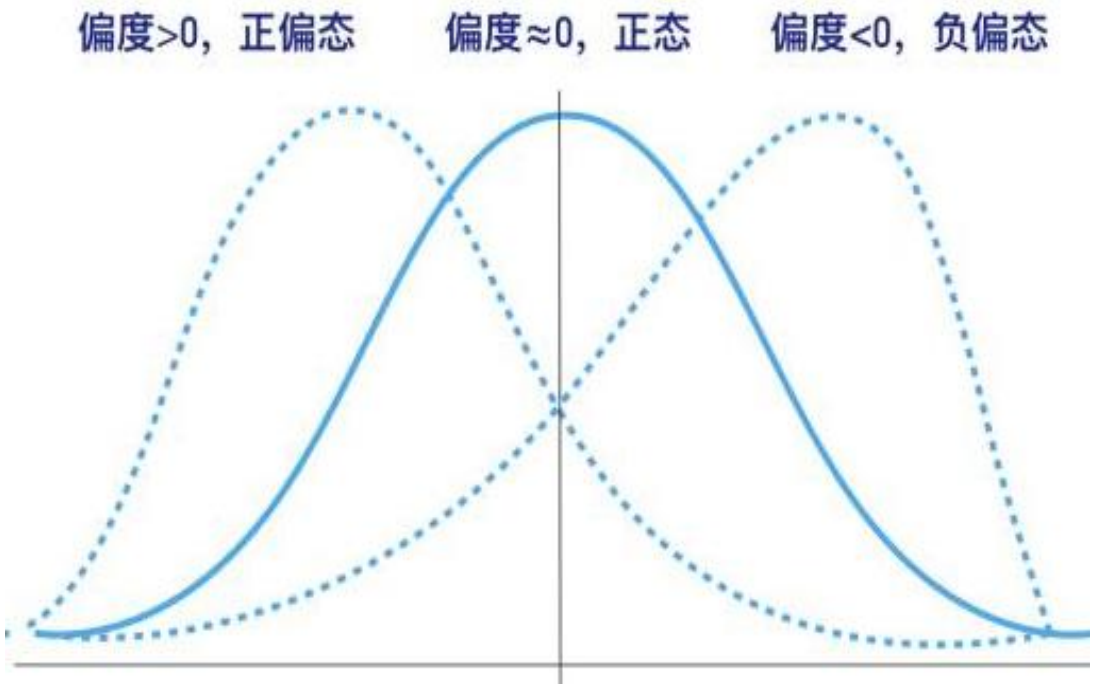
$$V_\sigma = \frac{\sigma}{\mu} \times 100\% \text{，或者 } V_s = \frac{s}{\bar{x}} \times 100\%$$

特点

- 不受标志值的计量单位和平均数大小的影响；是比较具有不同标准差和不同平均数的变量差异程度的一种较好统计量。
- 标准差系数比标准差能更好地刻画变量的差异程度。

2.3 数值法（续）

偏度（Skewness)



- 定义
  - 反映变量分布不对称方向和程度的
  - 种统计量。
- 公式

$$\alpha = \frac{M_3}{\sigma^3}$$

其中，$M_3 = \frac{1}{N}\sum_{i=1}^{N}(x_i - \mu)^3$ 为3阶中心矩。

- 判定标准 $\alpha \geq 0$
  - 当 $\alpha = 0$ 时，变量分布为正偏（或右偏）分布；
  - 当 $\alpha < 0$ 时，变量分布为对称分布；
  - 当　　时，变量分布为负偏（或左偏）分布。
  - 特别地，正态分布的偏度等于0，即正态分布是对称的。

2.3 数值法（续）

- 不同分布形态下平均数、众数与中位数之间的关系
  - 在正偏（右偏）分布下，平均数＞中位数＞众数；
  - 在对称分布下，平均数＝中位数＝众数；
  - 在负偏（左偏）分布下，平均数＜中位数＜众数。

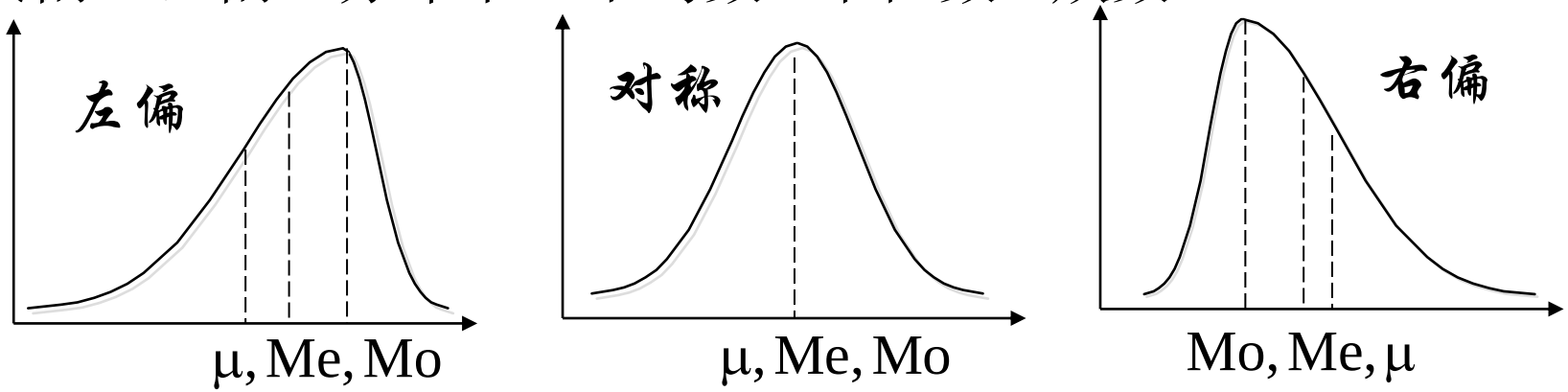


图2.3.2 均值、众数与中位数的关系

- 问题2.3.3
  - 当分布存在严重偏斜（正偏或负偏）时，平均数和中位数哪个更适合度量数据分布的集中趋势？

2.3 数值法（续）

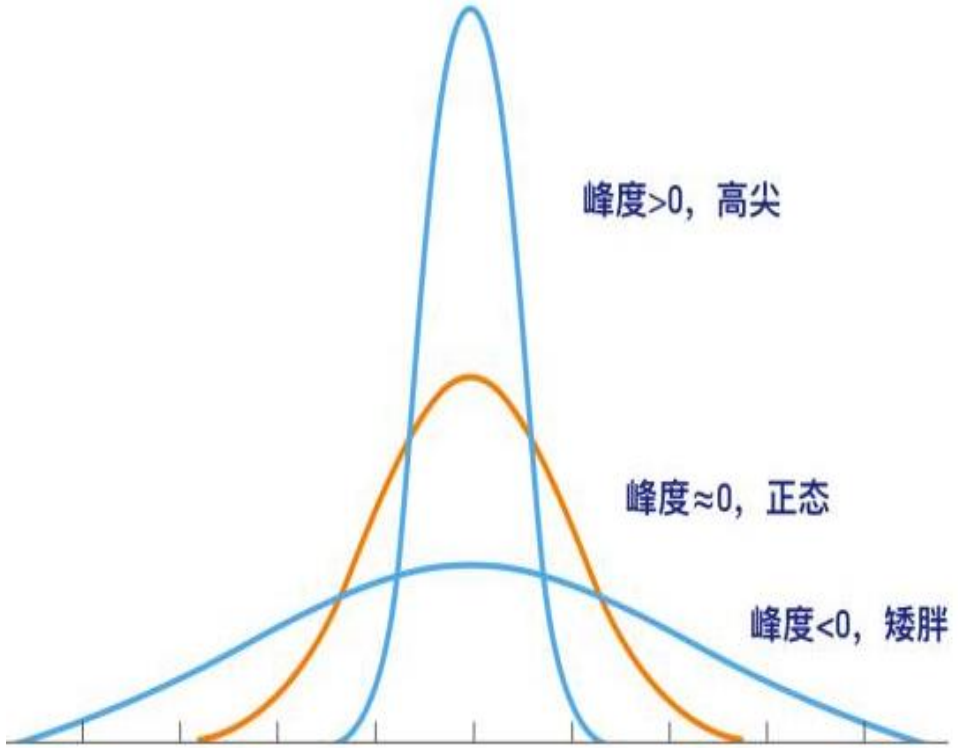
峰度（Kurtosis）

- 定义
  - 反映变量分布的尖峭或峰凸程度的一种统计量。
- 公式

$$\beta = \frac{M_4}{\sigma^4}$$

其中，$\mathrm{M}_4 = \frac{1}{N}\sum_{i=1}^{N}(x_i - \mu)^4$。

- 判定标准
  - 当 $\beta > 3$ 时，变量分布为高峰度；
  - 当 $\beta = 3$ 时，变量分布为正态峰度；
  - 当 $\beta < 3$ 时，变量分布为低峰度。
- 问题2.3.4
  - 高峰还是低峰分布的对比基准分布是哪种分布？



2.3 数值法（续）

探索性数据分析

五数概括法

- 五数包括：最小值、第1个四分位数、中位数、第3个四分位数、最大值。

箱形图（Box Plots）

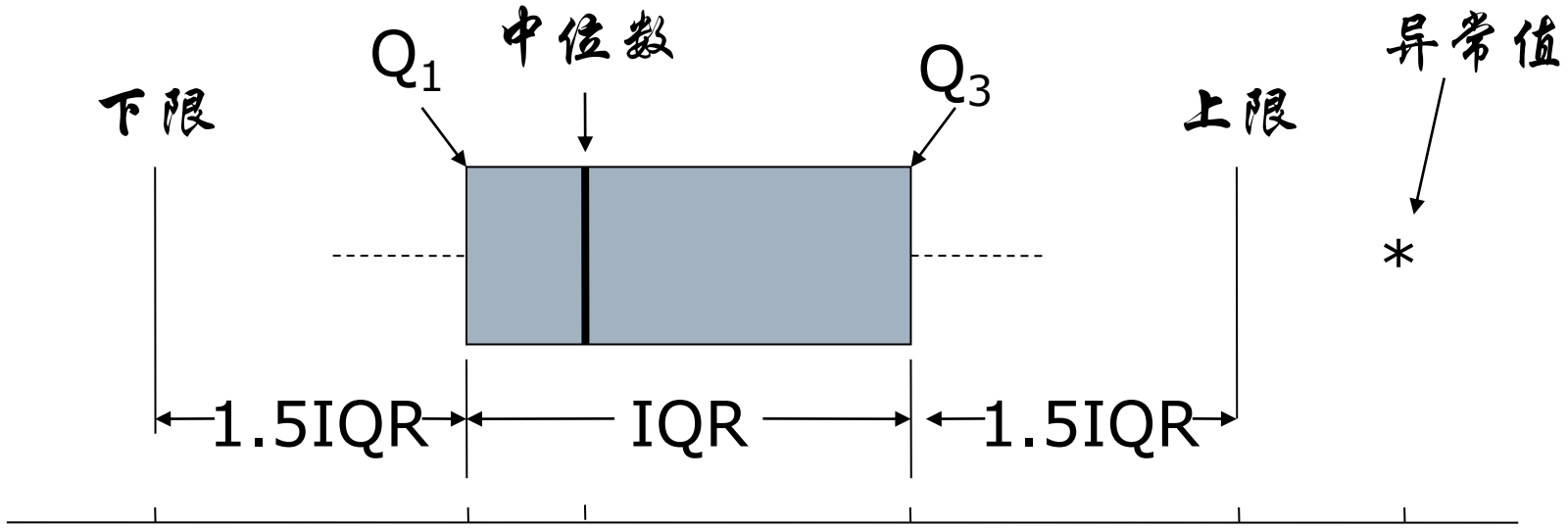
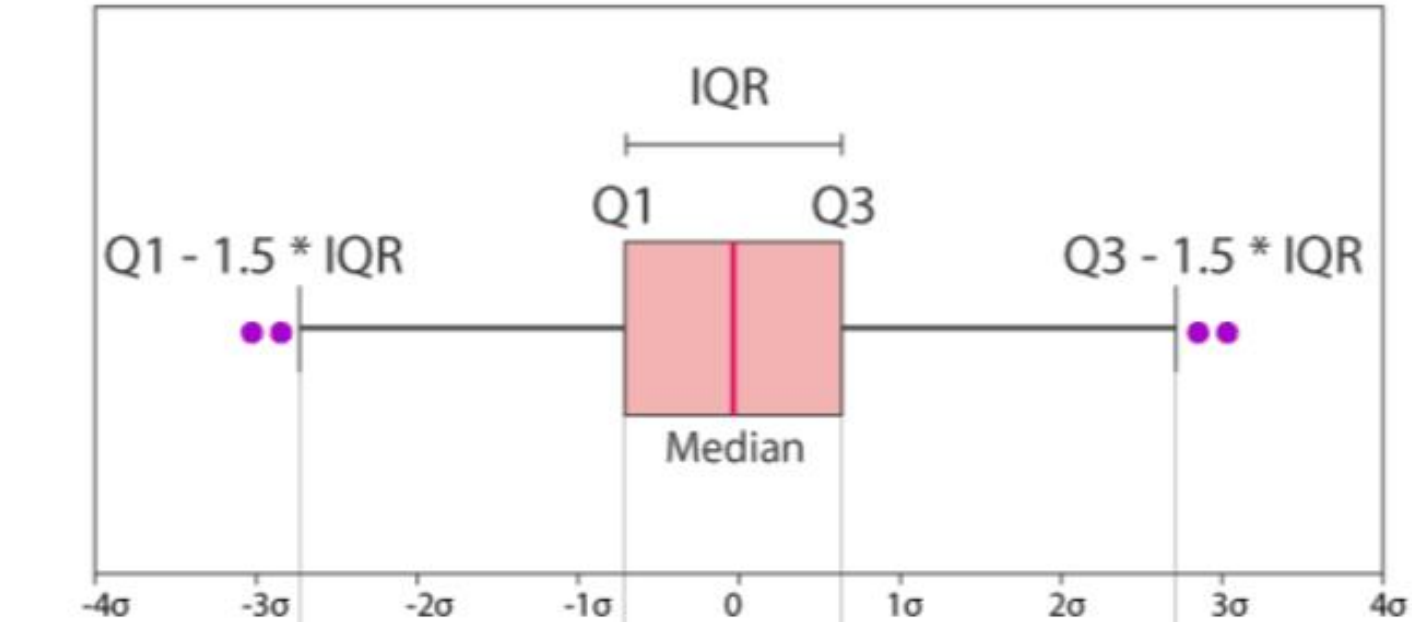
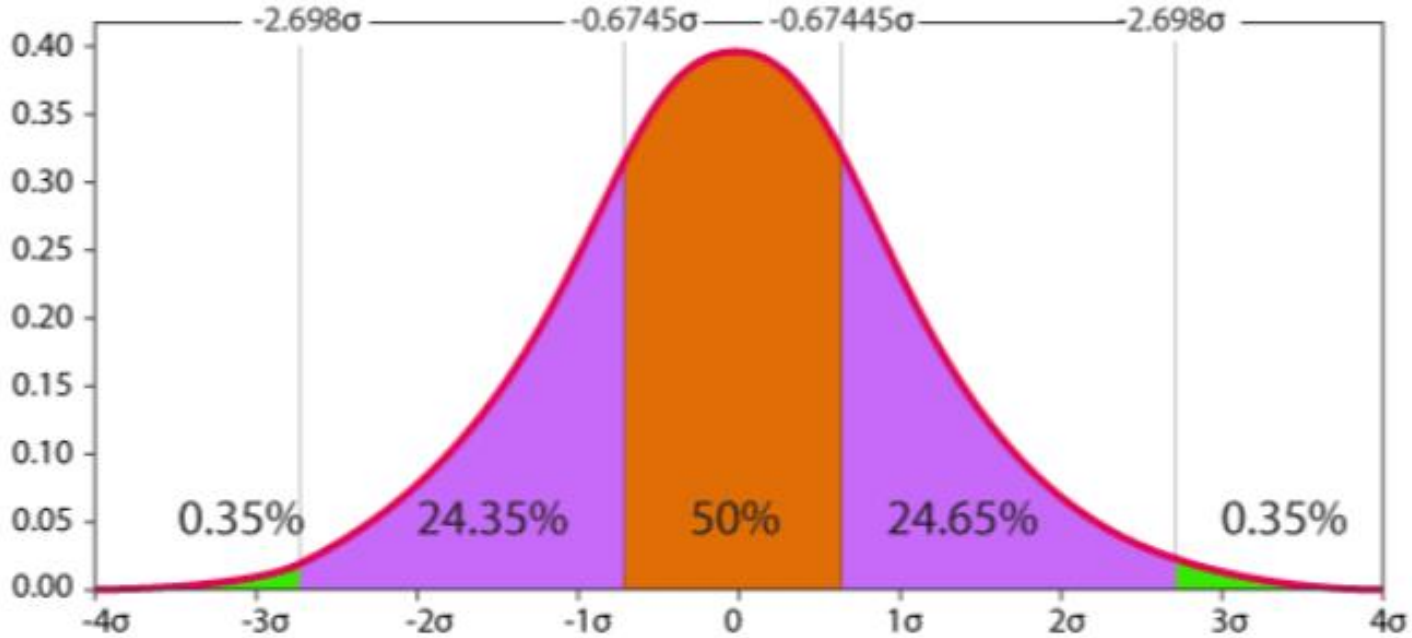


图2.3.4 箱形图


IQR
Q1
Q3
Q1 - 1.5 * IQR
Q3 - 1.5 * IQR
Median
-4σ
-3σ
-2σ
-1σ
0
1σ
2σ
3σ
4σ


-2.698σ
-0.6745σ
-0.67445σ
2.698σ
0.40
0.35
0.30
0.25
0.20
0.15
0.10
0.05
0.00
0.35%
24.35%
50%
24.65%
0.35%
-4σ
-3σ
-2σ
-1σ
0
1σ
2σ
3σ
4σ

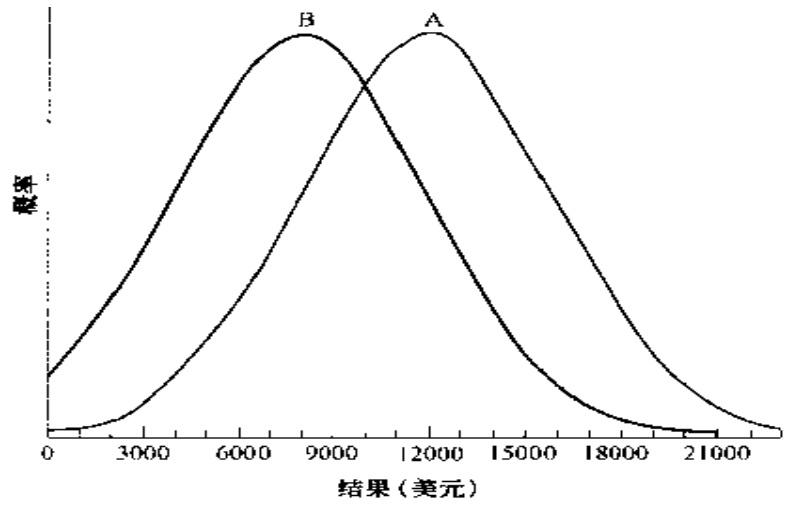
偏态判断方法

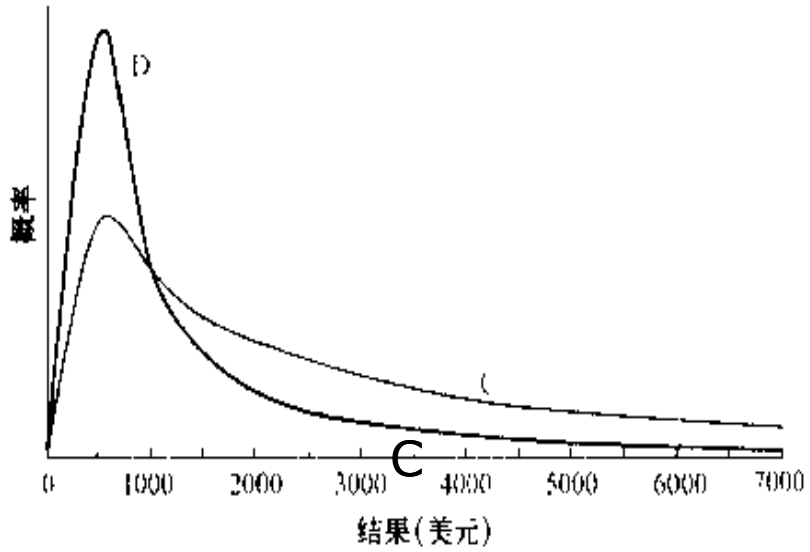
- 当中位数到上限的距离大于中位数到下限的距离时，则数据呈现正偏，否则为负偏。

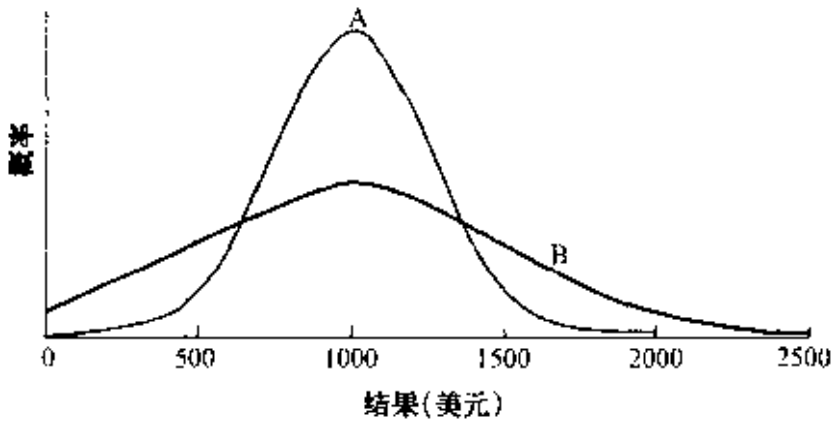
异常值判断方法

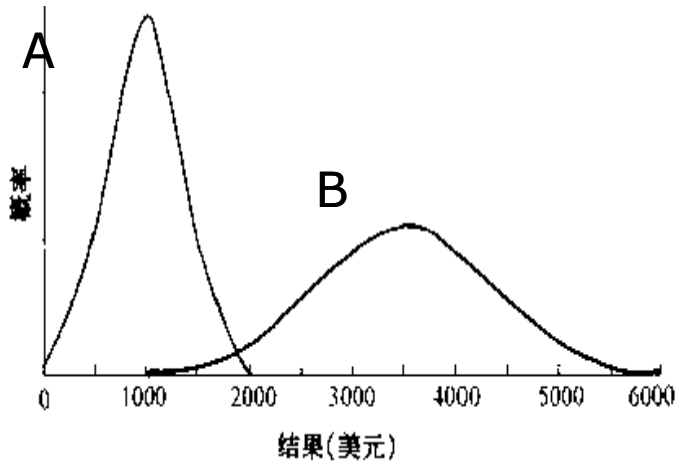
- 当数据值落在排序数据的上限右侧或下限左侧时，则该数据值为异常值。

- 比较概率分布的变量均值和标准差。

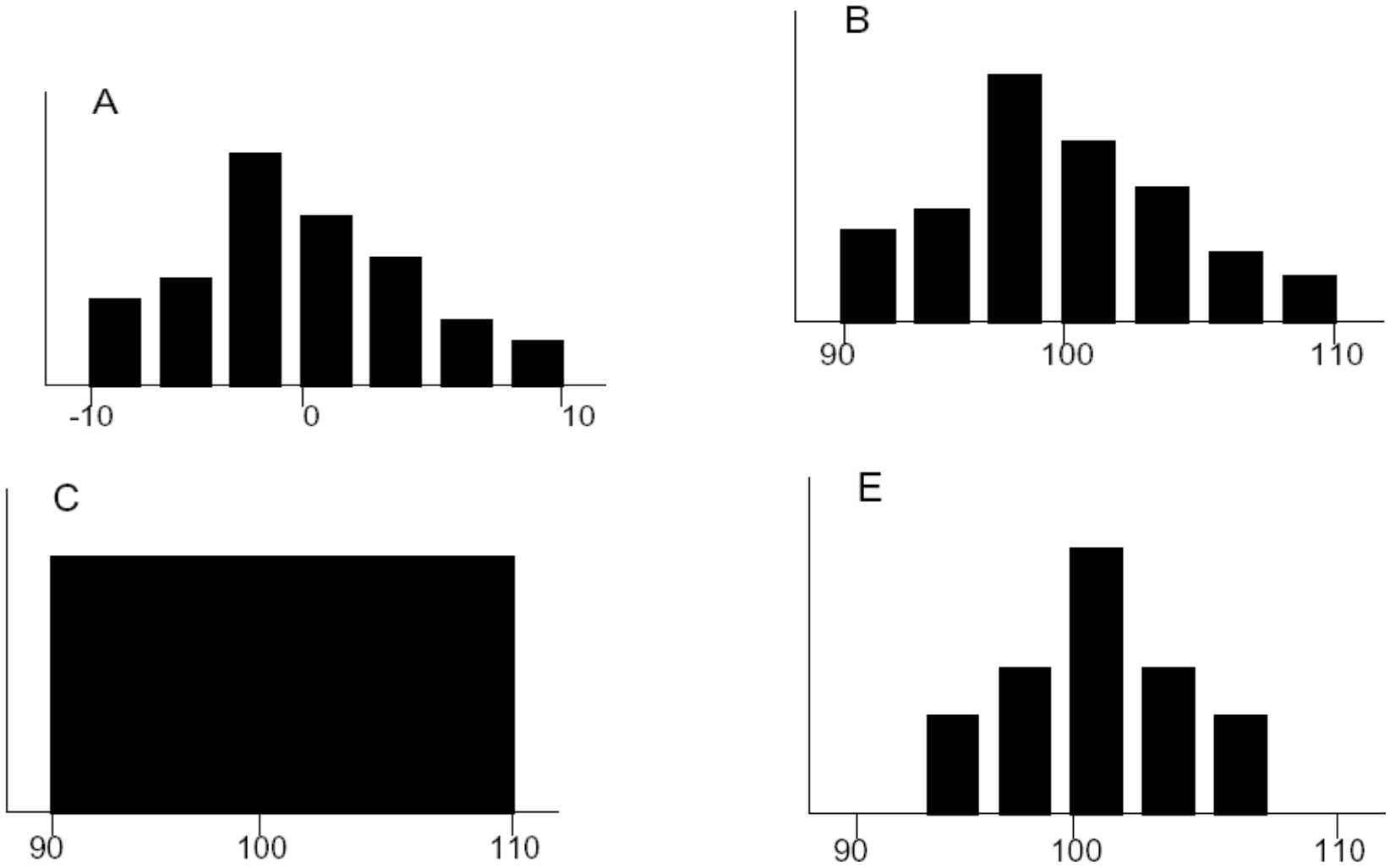








概率分布图



概率分布图2

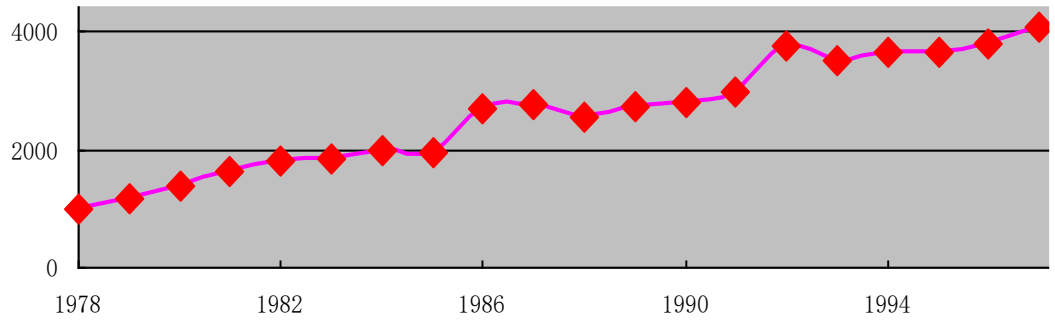
2.5 谨防统计“陷阱”

精心挑选的平均指标

- 比如，政府在设置税率时是选择平均数还是选择中位数？

惊人的统计图形

- 哪个图表示更快的增长速度？



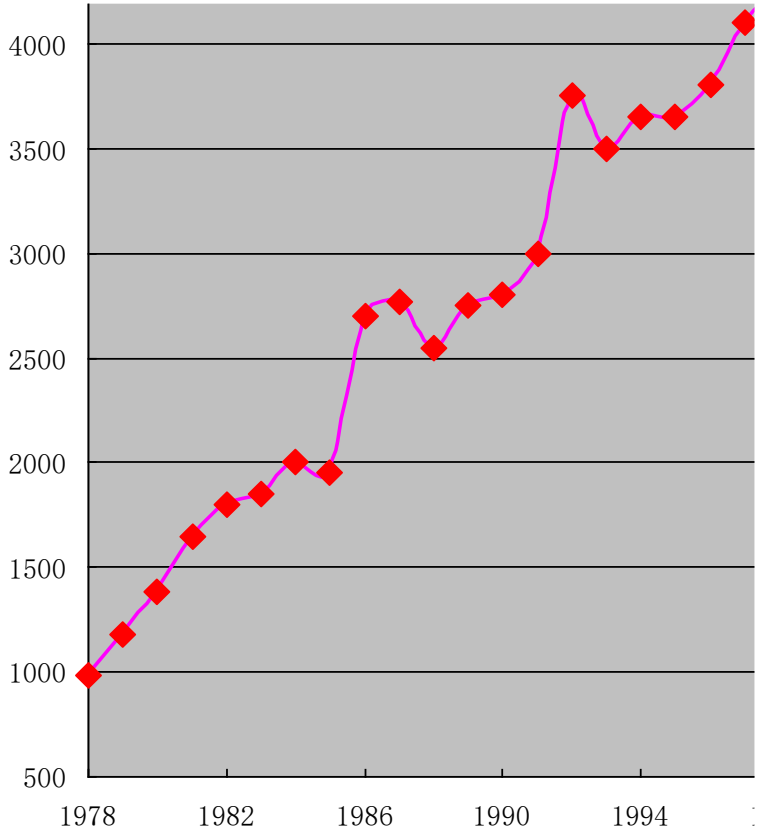
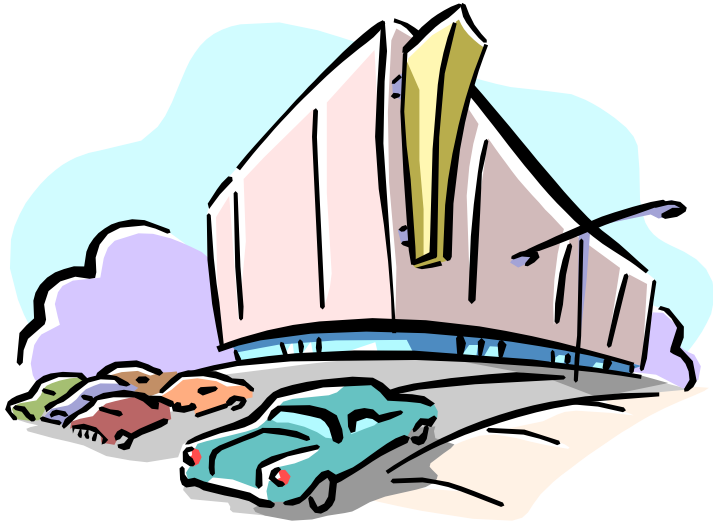


图2.5.1 图形的欺骗性

2.5 谨防统计“陷阱”（续）

防止对策

- 反问五个问题：
  - 谁说的
  - 他是如何知道的
  - 遗漏了什么
  - 是否有人偷换了概念
  - 这个资料有意义吗



简要回顾：描述统计—数值法

类型	名称		定义	计算公式/方法（以样本为例）	意义	图示	备注
平均水平	平均值						
	中位数						
	众数						
	百分位数						
	四分位数						
变异程度	极差						
	四分内距						
	方差						
	标准差						
	标准差系数						
分布形态	偏度						
	峰度						

描述统计的主要问题回顾

- 1. 描述统计如何认识随机现象的数量规律
- 2. 用于数据特征分析的主要指标及其特点
- 3. 众数、中位数和平均数的特点及联系
- 4. 分位数的计算步骤
- 5. 直方图和柱状图的绘制步骤

➢ 上证指数、深证成指和沪深300日收益的描述统计

表 4.7 我国主要股票指数日收益的描述统计分析（2011.1.1-2015.12.31）

	上证指数（000001.SH）	深证成指（399001.SZ）	沪深300（000300.SH）
个数	1214	1214	1214
最小值	-0.0887	-0.0860	-0.0915
最大值	0.0560	0.0625	0.6499
均值	0.000191	0.000014	0.000145
标准差	0.0150	0.0173	0.01617
偏度	-0.8587	-0.5493	-0.5960
峰度	8.6605	6.0296	7.4455