# 描述性统计分析

# 要点

01 统计的基本概念

02 数据的描述性统计

03 统计分布

04 相关分析

# 统计的基本概念

01

# 统计学含义

CDA数据分析师 CERTIFIED DATA ANALYST

统计学是一门收集、处理、分析、解释数据并从数据中得出结论的科学

**【数据分析步骤】 核心：数据**

收集数据 → 处理数据 → 分析数据 → 解释数据

分析数据：

- 描述性统计分析
- 推断性统计分析

# 统计学分析数据的方法

CDA数据分析师 CERTIFIED DATA ANALYST

**描述性分析** **研究数据收集、处理和描述的统计学方法**

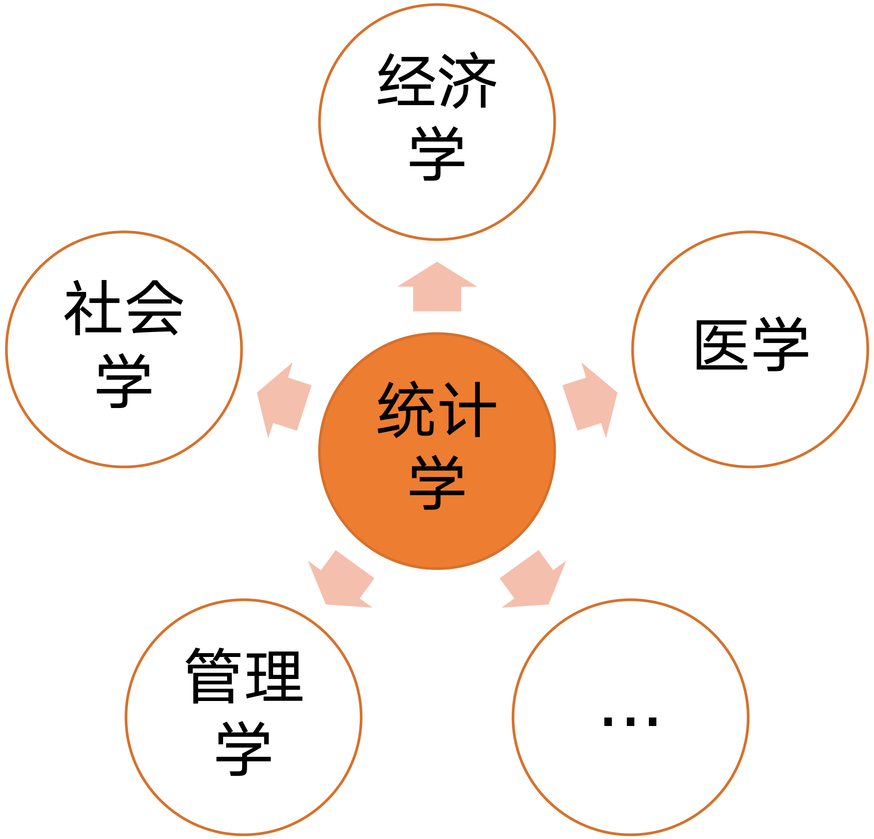
总体规模、对比关系、集中趋势、离散程度、偏态、峰态、……

**推断性分析** **研究如何利用样本数据来推断总体特征的统计学方法**

估计、假设检验、列联分析、方差分析、相关分析、回归分析、……

# 统计学应用

随着计算机的发展及各种统计软件的开发，作为一门基础学科的统计学在金融、保险、生物、经济等领域得到了广泛应用。



# 统计学的基本概念—数据

1、统计学的对象是**数据**。

2、数据的形式

- **数字**　可以进行比较、加减乘除等运算，严格的数据符号，常用阿拉伯数字表示
- **文字**　不可运算，如男、女等

【思考1】：阿拉伯数字一定是数字吗？

不一定。在处理数据时，我们有时候把男记作1，女记作0，此时的1和0不是数字。
实际上，阿拉伯数字只是一个代替的符号而已，阿拉伯数字符号也可以表示文字。

【思考2】：以下是数据吗？

1000　1个人　20岁　女同学　第一名　评价很好

# 数据的分类

1、按照计量尺度分类：

- 分类型数据：对事物进行分类的结果，如人的性别分为：男、女
- 顺序型数据：对事物类别顺序的测度，如产品分为：一等品、二等品、三等品
- 数值型数据：对事物的精确测度，如身高为：175cm、180cm

2、特点：

分类型数据：不可排序，不可计算

顺序型数据：可排序，不可计算

数值型数据：可排序，可计算

定性数据

定量数据

低级

高级

# 数据的分类

【思考】：已知某人月收入3000元，则：

"3000元"是什么类型数据？ 数值型

3000元属于区间[2000，4000]，则[2000，4000]是什么类型数据？ 数值型

[2000，4000]属于中等收入，则"中等"是什么类型数据？ 顺序型

补充：1、区间（分组的数值型数据）仍属于数值型

2、不同类型的数据之间可以进行转换，低级数据的方法高级数据可以用，但是高级数据的方法低级数据不可以用。

# 数据的其他分类

1、按来源不同：
- 直接来源（一手数据、原始资料）
- 间接来源（二手数据、次级资料）

2、按收集方式不同：
- 观测的数据
- 实验的数据

3、按与时间的关系不同：
- 截面数据
- 时间序列数据
- 混合数据（面板数据）

4、按概型不同：
- 离散型数据
- 连续型数据

5、一种特殊的数据：虚拟变量数据

# 统计学的基本概念—总体和样本

1、**总体(population)**
指研究的所有元素的集合。其中每个元素称为个体。
如：现研究全校学生的平均年龄，总体是：全校学生
和总体相关的事物，统计学上用希腊字母表示。

【思考】实际中，总体的个体往往难于一一研究，如何解决？
——抽取样本

2、**样本 (sample)**
从总体中抽取的一部分元素的集合。
如：为研究全校学生的平均年龄，由于总体太大，从中抽取100人进行研究，该研究中的样本是抽取的这100个学生。
和样本相关的事物，统计学上用英文字母表示。

构成样本的元素的数目称为**样本容量**。

# 统计学的基本概念—参数和统计量

**1、参数(parameter)**
指研究者想要了解的总体的某种特征值
主要有总体均值($\mu$)、总体标准差($\sigma$)、总体比例($\pi$)等

**2、统计量(statistic)**
指根据样本数据计算出来的一个量，即样本的某个特征值；
常见的统计量有样本均值($\bar{x}$)、样本标准差(s)、样本比例(p)等。

# 统计学的基本概念—变量

**1、变量**
指描述事物某种特征的概念。如商品销售额、受教育程度、产品的质量等级等。

**2、变量与数据的关系：**变量的具体表现称为变量值，即数据。

**3、变量的分类：**根据变量的数据计量尺度不同来分

分类变量(categorical variable)：说明事物类别的一个名称
顺序变量(rank variable)：说明事物有序类别的一个名称
数值型变量(metric variable)：说明事物数字特征的一个名称

# 数据的描述性统计

02

# 描述统计

思考：某超市后台记录了一年内53万余条消费者的消费数据，请问如何做描述统计分析？（撰写一份数据描述统计分析报告）

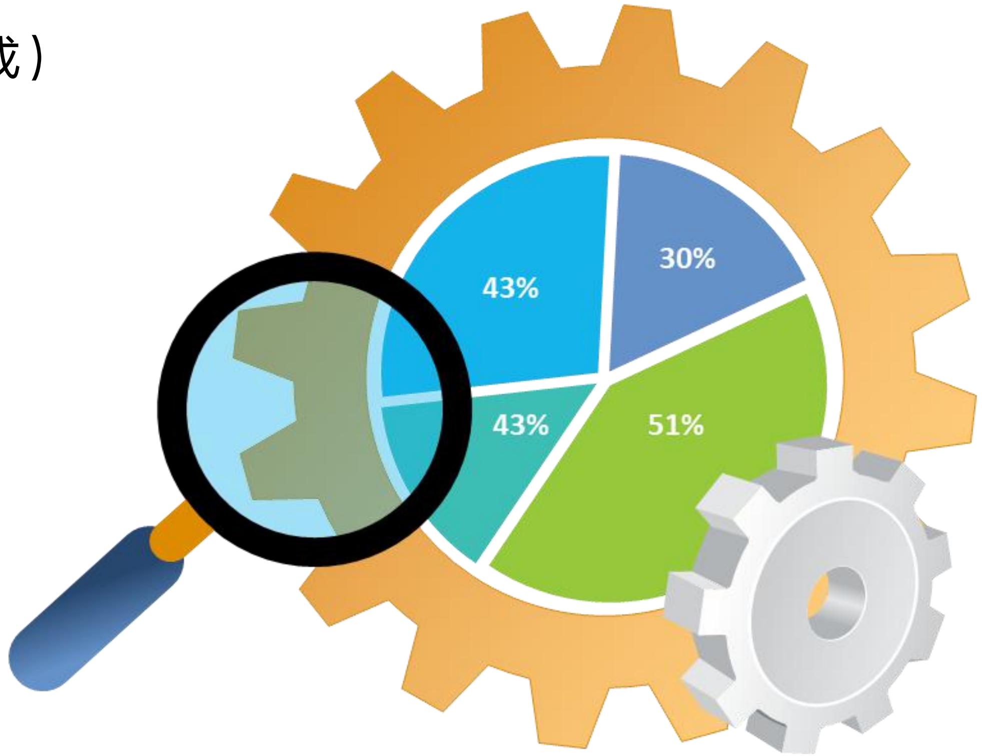
**五个角度：**

一、总体规模的描述——总量指标
二、对比关系的描述——相对指标
三、集中趋势的描述——平均指标
四、离散程度的描述——变异指标
五、分布形态的描述——偏态与峰态
六、描述性统计图表

# 总量指标、相对指标

**1.总量指标：**反映在一定时间、空间条件下某种现象的总体规模、总水平或总成果的统计指标。
如：营业额、利润

**2.相对指标：**是两个有相互联系的指标数值之比。
如：目标完成率（实际完成/计划完成）

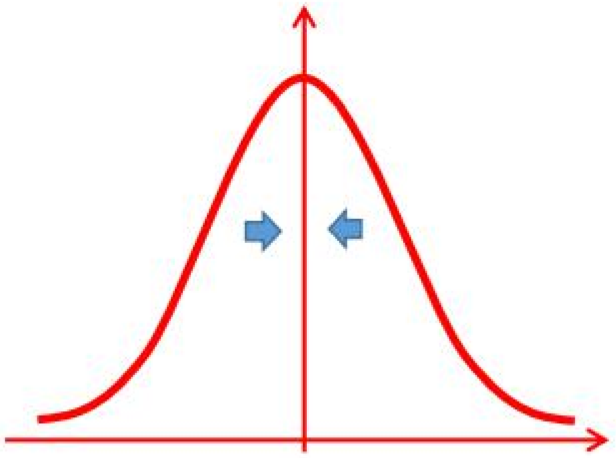


# 集中趋势的描述—平均指标

集中趋势(Central tendency)

1、定义：一组数据向其中心值靠拢的趋势
2、测度集中趋势就是寻找数据水平的代表值或中心值

分类型数据可用　众数
顺序型数据可用　众数、分位数
数值型数据可用　众数、分位数、均值



# 集中趋势的描述—平均指标

CDA数据分析师 CERTIFIED DATA ANALYST

## 一、众数

1、定义：出现次数最多的变量值
2、表示的符号：$M_0$
3、计算：寻找数据中出现次数最多的值

**众数的不唯一性**

【练习】：

| | |
|---|---|
| 7 5 9 17 6 8 | 无众数 |
| 4 5 7 8 5 5 | 1个众数 |
| 21 26 26 36 43 43 | 2个众数 |

# 集中趋势的描述—平均指标

## 二、分位数

是指根据对数据位置进行划分，处于某些特定位置上的数。常用的分位数有二分位数（也叫“中位数”）、四分位数、十分位数、百分位数等，课程重点介绍中位数和四分位数。

### （一）中位数 (二分位数)

1、定义：数据排序后，处于中间位置上的值

2、表示的符号：$M_e$

3、计算：数据的个数为$n$，则中位数的位置$=\frac{n+1}{2}$

# 集中趋势的描述—平均指标

中位数:

【练习1】9个家庭的人均月收入数据:

1500 745 781 1070 877 990 3000 1450 1680

排序: 745 781 877 990 1070 1450 1500 1680 3000

位置: 1 2 3 4 5 6 7 8 9

$$中位数位置 = \frac{n+1}{2} = \frac{9+1}{2} = 5$$

中位数: 1070

【练习2】10个家庭的人均月收入数据

853 798 770 630 966 2100 1930 1350 1600 1089

排序: 630 770 798 853 966 1089 1350 1600 1930 2100

位置: 1 2 3 4 5 6 7 8 9 10

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

$$中位数: \frac{966+1089}{2} = 1027.5$$

# 集中趋势的描述—平均指标

## （二）四分位数

1、定义：四分位数分为：下四分位数和上四分位数两种，指排序后处于25%和75%位置上的值

2、表示的符号：下四分位数$Q_L$ ，上四分位数$Q_U$

3、计算：数据的个数为$n$，则

下四分位数$Q_L$的位置：$\frac{n}{4}$

上四分位数$Q_U$的位置：$\frac{3n}{4}$

# 集中趋势的描述—平均指标

四分位数：

【练习1】12个家庭的人均月收入数据

原始数据:1500,750,780,1080,660,850,960,2000,1250,1630,530,2100

排序:530,660,750,780,850,960,1080,1250,1500,1630,2000,2100

位置: 1 2 3 4 5 6 7 8 9 10 11 12

$$Q_u\text{的位置} = \frac{3n}{4} = \frac{3\times 12}{4} = 9$$

$$Q_u = 1500$$

$$Q_l\text{的位置} = \frac{n}{4} = \frac{12}{4} = 3$$

$$Q_l = 750$$

【练习2】10个家庭的人均月收入数据

原始数据:1500,750,780,1080,660,850,960,2000,1250,1630

排序: 660,750,780,850,960,1080,1250,1500,1630,2000

位置: 1 2 3 4 5 6 7 8 9 10

$$Q_u\text{的位置} = \frac{3n}{4} = \frac{3\times 10}{4} = 7.5$$

$$Q_u = \frac{1250+1500}{2} = 1375$$

$$Q_l\text{的位置} = \frac{n}{4} = \frac{10}{4} = 2.5$$

$$Q_l = \frac{750+850}{2} = 800$$

# 集中趋势的描述—平均指标

## 三、均值(mean)

### （一）算术平均数

1、定义：数据的和与数据个数之比

2、表示的符号：$\bar{x}$

3、计算：

- 简单算术平均数（根据未分组数据计算的）：$\bar{x}=\dfrac{x_1+x_2+\cdots+x_n}{n}=\dfrac{\sum_{i=1}^{n}x_i}{n}$

- 加权算术平均数（根据分组数据计算的）：$\bar{x}=\dfrac{M_1f_1+M_2f_2+\cdots+M_kf_k}{f_1+f_2+\cdots+f_k}=\dfrac{\sum_{i=1}^{k}M_if_i}{n}$

（其中：数据个数为n，分组数据的组数为k， $M_i$为组中值， $f_i$为各组的频数。）

4、特点：易受极端值影响

# 加权算术平均数

CDA数据分析师 CERTIFIED DATA ANALYST

| 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 |

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

# 集中趋势的描述—平均指标

CDA数据分析师 CERTIFIED DATA ANALYST

（二）几何平均数

1、定义：n个变量值乘积的n次方根

2、表示的符号：G

3、计算：

- 简单调和平均数（根据未分组数据计算的）：$G=\sqrt[n]{x_1 x_2 \ldots x_n}$

- 加权调和平均数（根据分组数据计算的）：$G=\sqrt[(f_1+f_2+\ldots+f_n)]{x_1^{f_1} x_2^{f_2} \ldots x_n^{f_n}}$

（其中：数据个数为n，分组数据的组数为k， $M_i$为组中值， $f_i$为各组的频数。）

4、特点：

a.易受极端值影响

b.常用于增长率数据的研究

c.所有数据需大于0

# 几何平均数

【例】一位投资者购持有一种股票，连续4年收益率分别为4.5%、2.1%、25.5%、1.9%。计算该投资者在这四年内的平均收益率

**几何平均：**

$$\bar{G} = \sqrt[4]{104.5\% \times 102.1\% \times 125.5\% \times 101.9\%} - 1$$

$$= 8.0787\%$$

**算术平均：**

$$\bar{G} = (4.5\% + 2.1\% + 25.5\% + 1.9\%) \div 4 = 8.5\%$$

# 集中趋势的描述—平均指标

（三）调和平均数

1、定义：变量值倒数的算数平均数的倒数

2、表示的符号：$H$

3、计算：

- 简单调和平均数（根据未分组数据计算的）：$H=\dfrac{n}{\dfrac{1}{x1}+\dfrac{1}{x2}+\cdots+\dfrac{1}{xn}}$

- 加权调和平均数（根据分组数据计算的）：$H=\dfrac{f_1+f_2+\cdots+f_k}{\dfrac{f_1}{M_1}+\dfrac{f_2}{M_2}+\cdots+\dfrac{f_k}{M_k}}$

（其中：数据个数为n，分组数据的组数为k， $M_i$为组中值， $f_i$ 为各组的频数。）

4、特点：

a.易受极端值影响
b.常用于效率数据的研究
c.有一项为0就无法计算H

# 集中趋势的描述—平均指标

CDA数据分析师 CERTIFIED DATA ANALYST

均值不等式

对于同一组数据，一定满足：

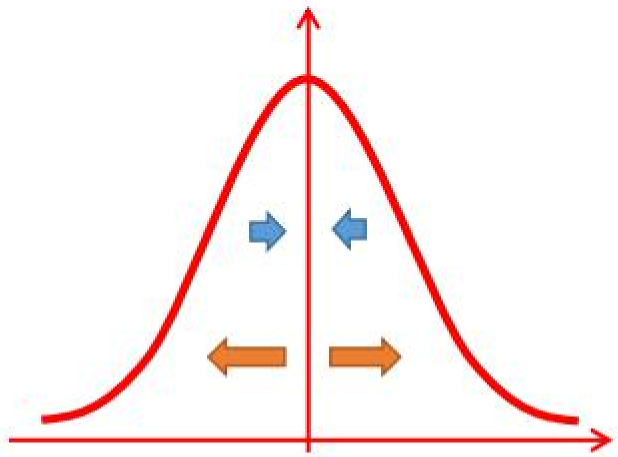
算术平均数 ≥ 几何平均数 ≥ 调和平均数

当所有数据取值相同的时候，等号成立。

# 离散程度的描述—变异指标

离散程度:

1、定义：反映各变量值远离其中心值的程度，是数据分布的另一个重要特征
2、从另一个侧面说明了集中趋势测度值的代表程度



# 离散程度的描述—变异指标

## 一、极差（range）

1、定义：一组数据的最大值与最小值之差
2、表示的符号： R
3、计算：R = max(xi) - min(xi)
4、特点：
    a.离散程度的最简单测度值
    b.极易受极端值影响
    c.未考虑数据的分布

【练习】计算数据：660 750 1500 2000的极差。

2000-660=1340

# 离散程度的描述—变异指标

## 二、平均差(mean deviation)

1、定义：各变量值与其均值离差绝对值的平均数；
2、表示的符号：$M_d$
3、计算：

未分组数据：$$M_d = \frac{\sum_{i=1}^{n} |x_i - \bar{x}|}{n}$$

分组数据：$$M_d = \frac{\sum_{i=1}^{k} |M_i - \bar{x}| f_i}{n}$$ （$M_i$为组中值）

4、特点：
a.能全面反映一组数据的离散程度：$M_d$越大，表示数据越分散。
b.数学性质较差，实际中应用较少

# 离散程度的描述—变异指标

【练习】计算2,4,7,10,10,10,12,12,14,15的平均差

| | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| 2 | 4 | 7 | 10 | 10 | 10 | 12 | 12 | 14 | 15 |
| -7.6 | -5.6 | -2.6 | 0.4 | 0.4 | 0.4 | 2.4 | 2.4 | 4.4 | 5.4 |
| 7.6 | 5.6 | 2.6 | 0.4 | 0.4 | 0.4 | 2.4 | 2.4 | 4.4 | 5.4 |

$\bar{x} = 9.6$

$$\frac{31.6}{10} = 3.16$$

# 离散程度的描述—变异指标

## 三、方差和标准差

（一）根据总体数据计算的，称为总体方差、总体标准差；
（二）根据样本数据计算的，称为样本方差、样本标准差；

1、定义：变量值与其算术平均数的离差的平方的算术平均数；
2、表示的符号：

总体方差：$\sigma^2$
总体标准差：$\sigma$
样本方差：$s^2$
样本标准差：$s$

3、特点：
a.数据离散程度的最常用测度值
b.反映了各变量值与均值的平均差异：方差或标准差越大，表示变量值与均值的平均差异越大

# 离散程度的描述—变异指标

4、计算：

总体方差：

未分组数据：$\sigma^2 = \frac{\sum_{i=1}^{N}(x_i - \mu)^2}{N}$

分组数据：$\sigma^2 = \frac{\sum_{i=1}^{K}(M_i - \mu)^2 f_i}{N}$ （$M_i$为组中值）

总体标准差：

$$\sigma = \sqrt{\sigma^2}$$

# 离散程度的描述—变异指标

4、计算:

样本方差:

未分组数据: $s^2 = \dfrac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n-1}$

分组数据: $s^2 = \dfrac{\sum_{i=1}^{k}(M_i - \bar{x})^2 f_i}{n-1}$ （$M_i$为组中值）

样本标准差:

$$s = \sqrt{s^2}$$

注: 样本方差计算公式的分母是n-1

- 样本方差自由度(degree of freedom)
- 自由度是指一组数据中可以自由取值的数据的个数

# 离散程度的描述—变异指标

## 四、离散系数（变异系数）

思考：哪组数据离散程度大？

元　　1，　2，　3

$$x_1=\frac{1+2+3}{3}=2,s_1^2=\frac{(1-2)^2+(2-2)^2+(3-2)^2}{3-1}=1,s_1=1$$

角　　10，20，30

$$x_2=\frac{10+20+30}{3}=20,s_2^2=\frac{(10-20)^2+(20-20)^2+(30-20)^2}{3-1}=100,s_2=10$$

因此，我们对标准差进行改进，得到离散系数。

# 离散程度的描述—变异指标

1、定义：是标准差与均值之比。
2、表示的符号：$V_s$
3、计算：$V_s = \frac{s}{\bar{x}}$
4、特点：
   a.是对数据相对离散程度的测度；
   b.消除了数据水平不同和数据计量单位不同对数据离散程度的影响；
   c.常用于对不同组别数据离散程度的比较。

# 相对位置的度量—标准化值

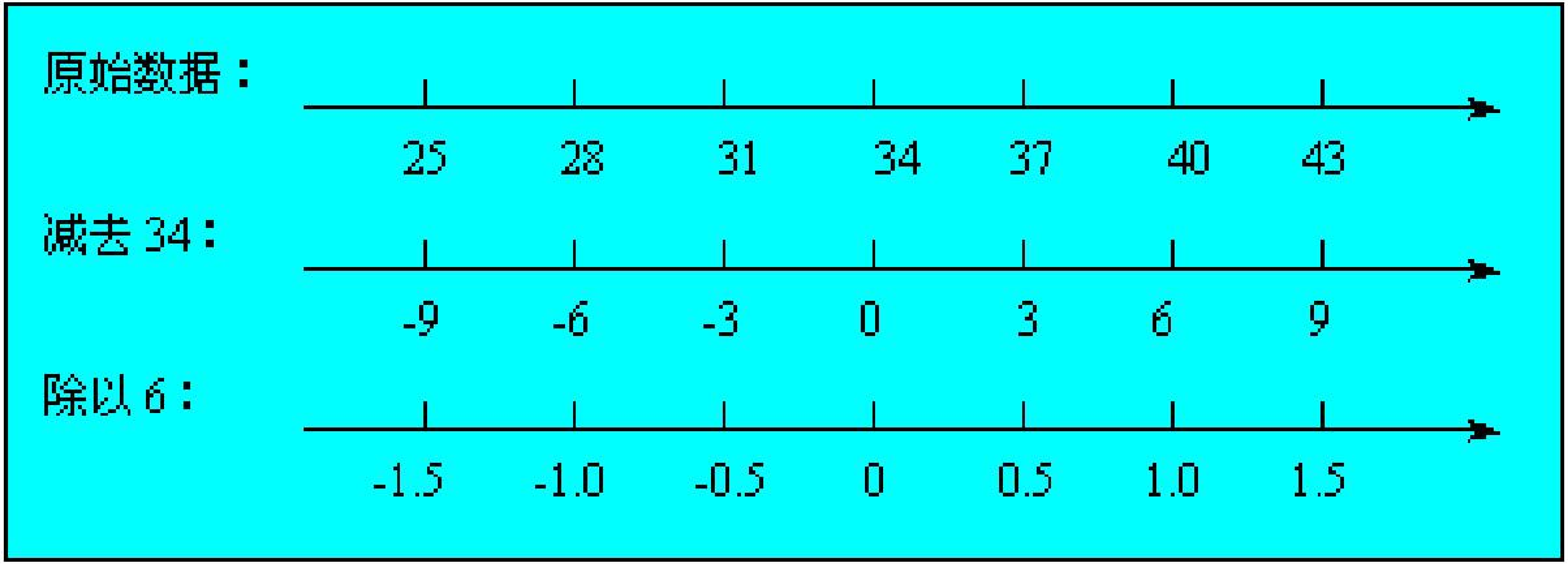
## 一、标准化值

1. 也称标准分数
2. 对某一个数据在全体中相对位置的度量
3. 可用于判断一组数据是否有离群值
4. 用于对变量的标准化处理
5. 计算公式为

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

# 相对位置的度量—标准化值

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



# 相对位置的度量—标准化值

- ➡经验法则表明：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

1. 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
2. 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”
3. 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

- ➡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

# 分布形态的描述—偏态与峰态

## 一、偏态(skewness)

1、定义：是指数据分布偏斜程度。

2、测量方法：使用偏态系数来测度数据的偏态。偏态系数用符号SK表示。

3、偏态系数的计算：（公式有多种，这里选常见的一种）

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

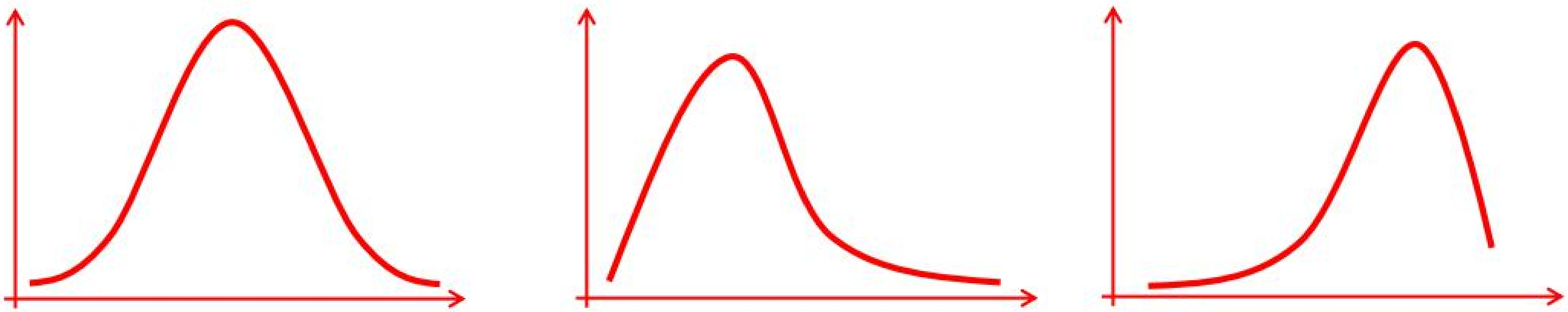
分组数据：$$SK=\frac{\sum(M_i-\bar{x})^3 f_i}{ns^3}$$

# 分布形态的描述—偏态与峰态

CDA数据分析师 CERTIFIED DATA ANALYST

4、偏态的判断：

（1）SK=0对称分布；　　SK>0右偏分布；　　SK<0左偏分布



（2）偏态的程度：

低度偏态分布：　$0 < |SK| \leq 0.5$

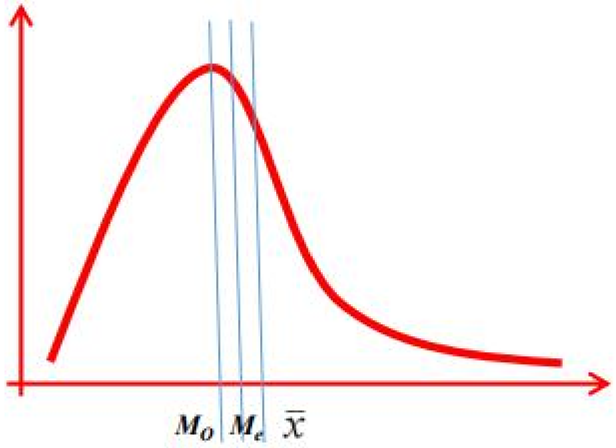
中等偏态分布：　$0.5 < |SK| \leq 1$

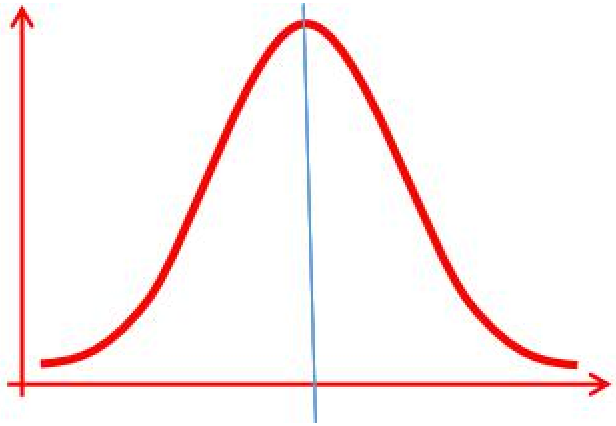
高度偏态分布：　$|SK| > 1$

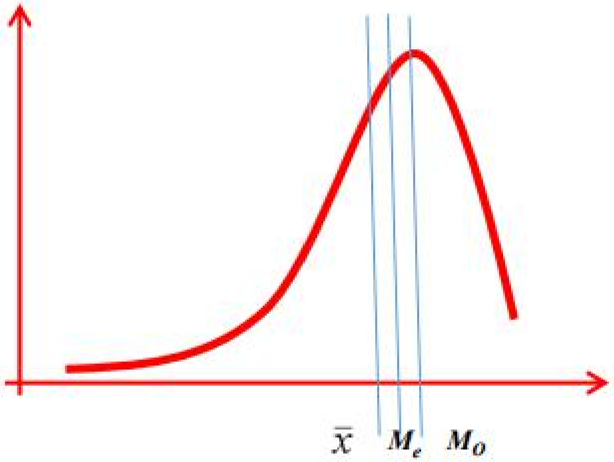
# 分布形态的描述—偏态与峰态



5、偏态对众数、中位数和均值之间关系的影响
对称分布：均值=中位数=众数
左偏分布：均值<中位数<众数
右偏分布：众数<中位数<均值





# 分布形态的描述—偏态与峰态

## 二、峰态(kurtosis)

1、定义：是指数据分布的扁平程度。

2、测量方法：使用峰态系数来测度数据的峰态。峰态系数用符号K表示。

3、峰态系数的计算：（公式有多种，这里选常见的一种）

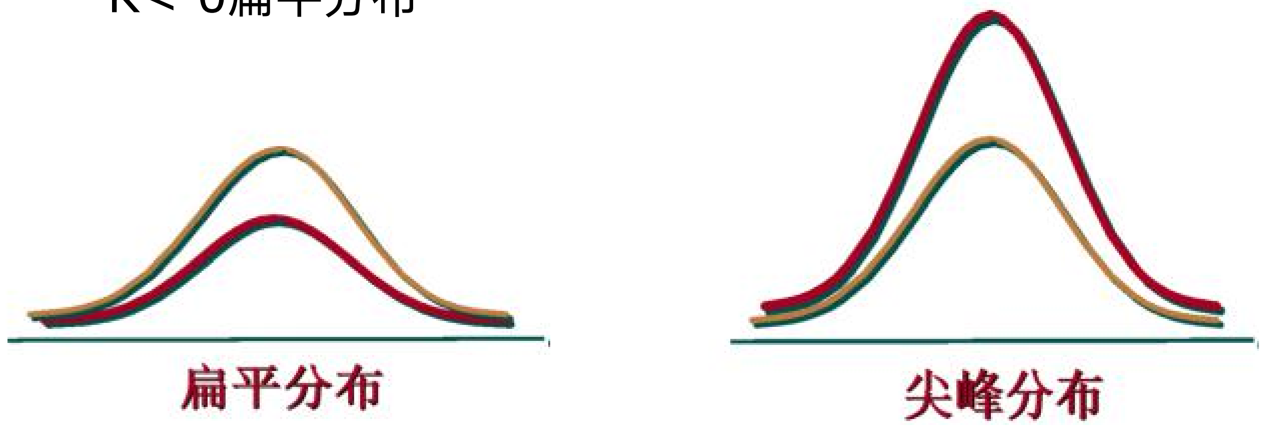
未分组数据：$$K=\frac{n(n+1)\sum(x_i-\bar{x})^4-3\left[\sum(x_i-\bar{x})^2\right]^2(n-1)}{(n-1)(n-2)(n-3)s^4}$$

分组数据：$$K=\frac{\sum(M_i-\bar{x})^4 f_i}{ns^4}-3$$

# 分布形态的描述—偏态与峰态

4、峰态的判断：

K=0扁平峰度适中
K> 0尖峰分布
K< 0扁平分布




5、峰态的程度：

低度尖峰分布：$0 < |K| \le 0.5$
中等尖峰分布：$0.5 < |K| \le 1$
高度尖峰分布：$|K| > 1$

# 描述性统计图表—直方图

**直方图**

定义：由一系列高度不等的矩形表示数据分布的情况。

**频数分布直方图**

1、定义：在统计数据时，横轴按组距分类，纵轴表示频数，每个矩形的高代表对应组距里数据的频数，称这样的统计图为频数分布直方图。

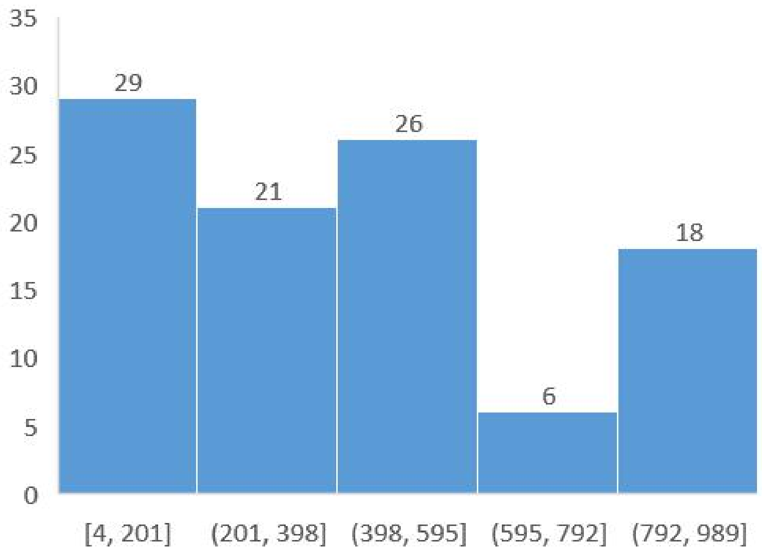
组数：把数据按照不同的范围分成几个组，分成的组的个数称为组数。

组距：每一组数据的极差。

2、特点：

a.能够显示各组频数分布的情况

b.易于显示各组之间频数的差别



# 描述性统计图表—直方图

**绘制直方图：**

1、收集数据。作直方图的数据一般大于50个
2、选择数据列，插入图表：直方图
3、确定组数、极差、组距

**绘制注意事项：**

1、抽取的样本数量过小，将会产生较大误差，可信度低，也就失去了统计的意义。因此，样本数不应少于50个。
2、组数选用不当，偏大或偏小，都会造成对分布状态的判断有误。

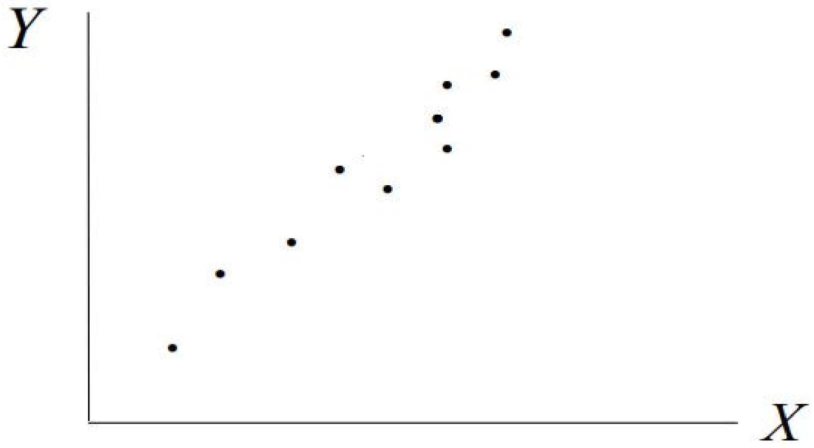
# 描述性统计图表—散点图

## 散点图

1、定义：数理统计分析中，数据点在平面直角坐标系上的分布图，表示因变量随自变量而变化的大致趋势。

2、特点：

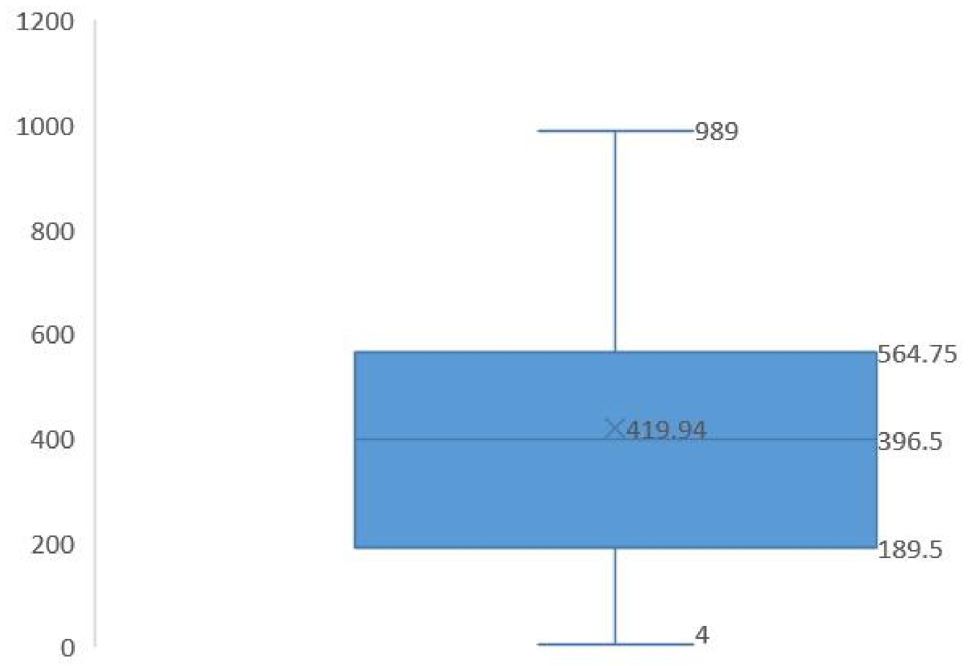
a.展示数据的分布情况
b.发现变量之间的关系



# 描述性统计图表—箱型图

**箱形图**

又称为盒须图或箱线图，显示一组数据分散情况的统计图




最大值

上四分位数

中位数　X：平均值

下四分位数

最小值

以Excel为例

# 统计分布

03

# 常用的分布

一、两点分布与二项分布 → 离散分布

二、正态分布
三、标准正态分布

连续分布

四、$x^2$分布
五、t分布
六、F分布

抽样分布

# 两点分布与二项分布

## 一、二项分布 X~B(n,p)

- E(X)=np
- D(X)=np(1-p)

| $X$ | 0 | 1 | 2 | … $k$ … | n |
|---|---|---|---|---|---|
| 概率 | $C_n^0 p^0 (1-p)^n$ | $C_n^1 p^1 (1-p)^{n-1}$ | $C_n^2 p^2 (1-p)^{n-2}$ | $C_n^k p^k (1-p)^{n-k}$ | $C_n^n p^n (1-p)^0$ |

## 二、两点分布 X~B(1,p)

n=1时的二项分布，又称伯努利分布

- E(X)=p
- D(X)=p(1-p)

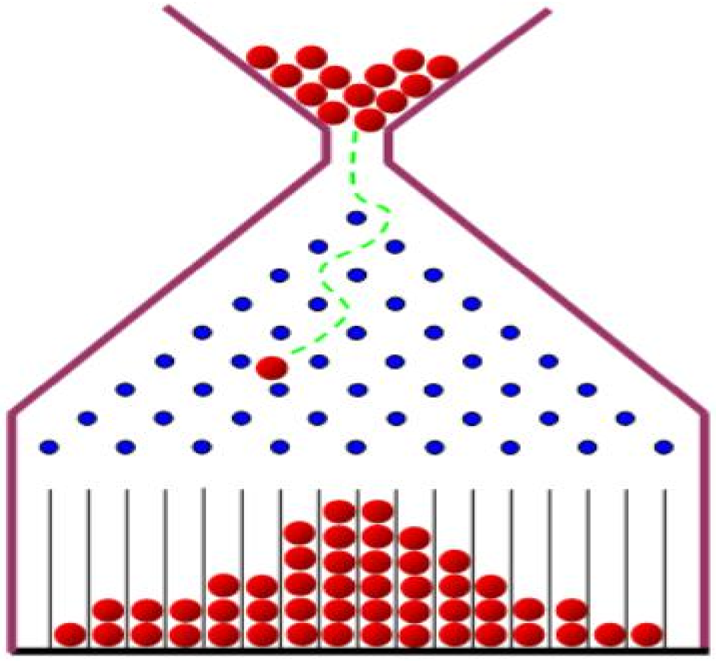
| $X$ | 0 | 1 |
|---|---|---|
| 概率 | 1-$p$ | $p$ |

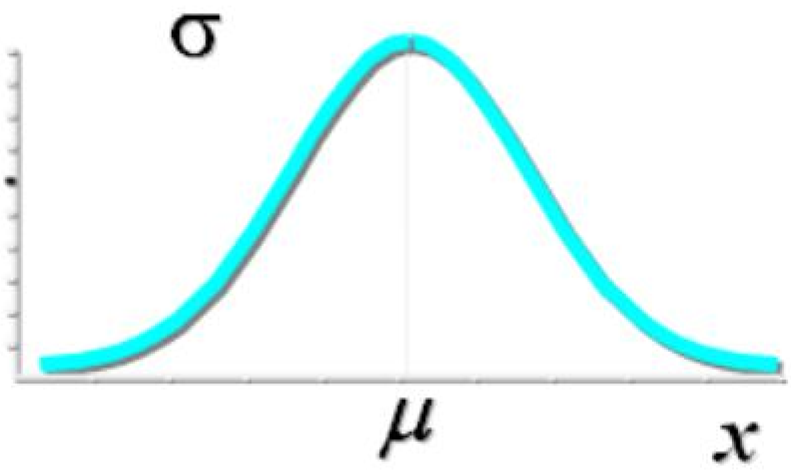
# 正态分布

1、表示方式：$X \sim N(\mu, \sigma^2)$

其中参数 $-\infty < \mu < \infty$ ，$\sigma > 0$

2、图像：






3、密度函数：

$$f(x)=\frac{1}{\sigma\sqrt{2\pi}}e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$

其中：

- f(x) = 随机变量 X 的频数
- $\sigma^2$= 总体方差
- $\pi \approx 3.14159$，$e \approx 2.71828$
- $x$= 随机变量的取值($-\infty < x < \infty$)
- $\mu$= 总体均值

# 正态分布

4、正态分布函数和图像的性质：

- 概率密度函数在x轴的上方，即f (x)>0；
- 正态曲线的最高点在均值$\mu$，它也是分布的中位数和众数；
- 正态分布是一个分布族，每一特定正态分布通过均值$\mu$的标准差$\sigma$来区分；

  $\mu$决定曲线的高度的位置，$\sigma$决定曲线的平缓程度，即宽度；
- 曲线 f(x)相对于均值$\mu$对称，尾端向两个方向无限延伸，且理论上永远不会与横轴相交；
- 正态曲线下的总面积等于1；

# 正态分布

5、正态分布的分布函数：

$$F(x_\circ)=P(x\le x_\circ)=\int_{-\infty}^{x_\circ} f(x)\mathrm{d}x=\int_{-\infty}^{x_\circ}\frac{1}{\sigma\sqrt{2\pi}}e^{-\frac{(x-\mu)^2}{2\sigma^2}}dx$$

6、若随机变量$X\sim N(\mu,\sigma^2)$，则：

$$P(X\le c)=\emptyset(\frac{c-\mu}{\sigma})$$

$$P(a<X\le b)=\emptyset(\frac{b-\mu}{\sigma})-\emptyset(\frac{a-\mu}{\sigma})$$

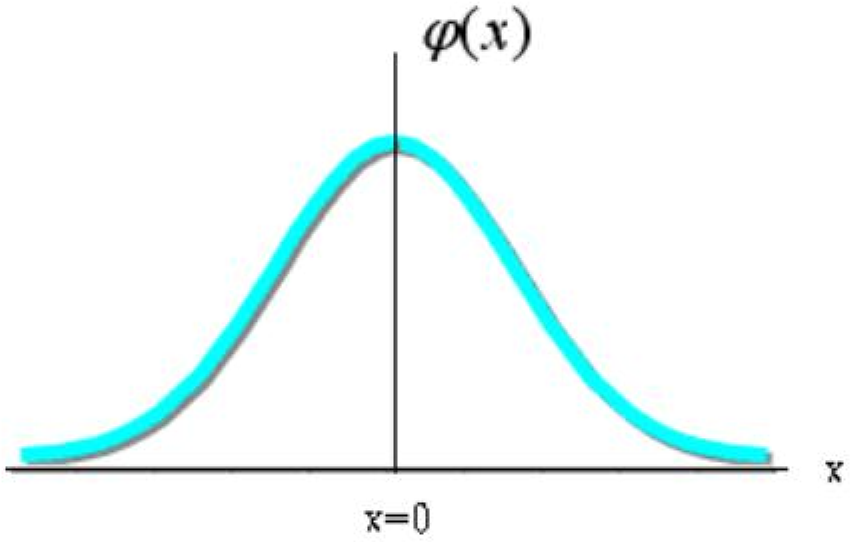
$$P(X>a)=1-\emptyset(\frac{a-\mu}{\sigma})$$

# 标准正态分布

1、表示方式：X~N(0,1)

称$\mu = 0, \sigma = 1$时的正态分布N(0,1)为标准正态分布

2、图像：




3、概率密度函数：

$$\varphi(x) = \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}} \quad , \quad -\infty < x < \infty$$

# 标准正态分布

4、标准正态分布函数和图象的性质

- 概率密度函数在x轴的上方，即φ (x)>0；
- 标准正态曲线的最高点在均值0，它也是分布的中位数和众数；
- 标准正态分布的曲线是唯一的，确定的，其高度和宽度是确定的。
- 曲线 φ(x)相对于均值x=0对称，尾端向两个方向无限延伸，且理论上永远不会与横轴相交；
- 标准正态曲线下的总面积等于1；

5、标准正态分布的分布函数：

$$\Phi(x_{\circ}) = p(x \le x_{\circ}) = \int_{-\infty}^{x_{\circ}} \varphi(x)\mathrm{d}t = \int_{-\infty}^{x_{\circ}} \frac{1}{\sqrt{2\pi}} \mathrm{e}^{-\frac{x^2}{2}} \mathrm{d}x$$

# 标准正态分布

6、设：$X \sim N(0,1)$，$a, b > 0$，则：

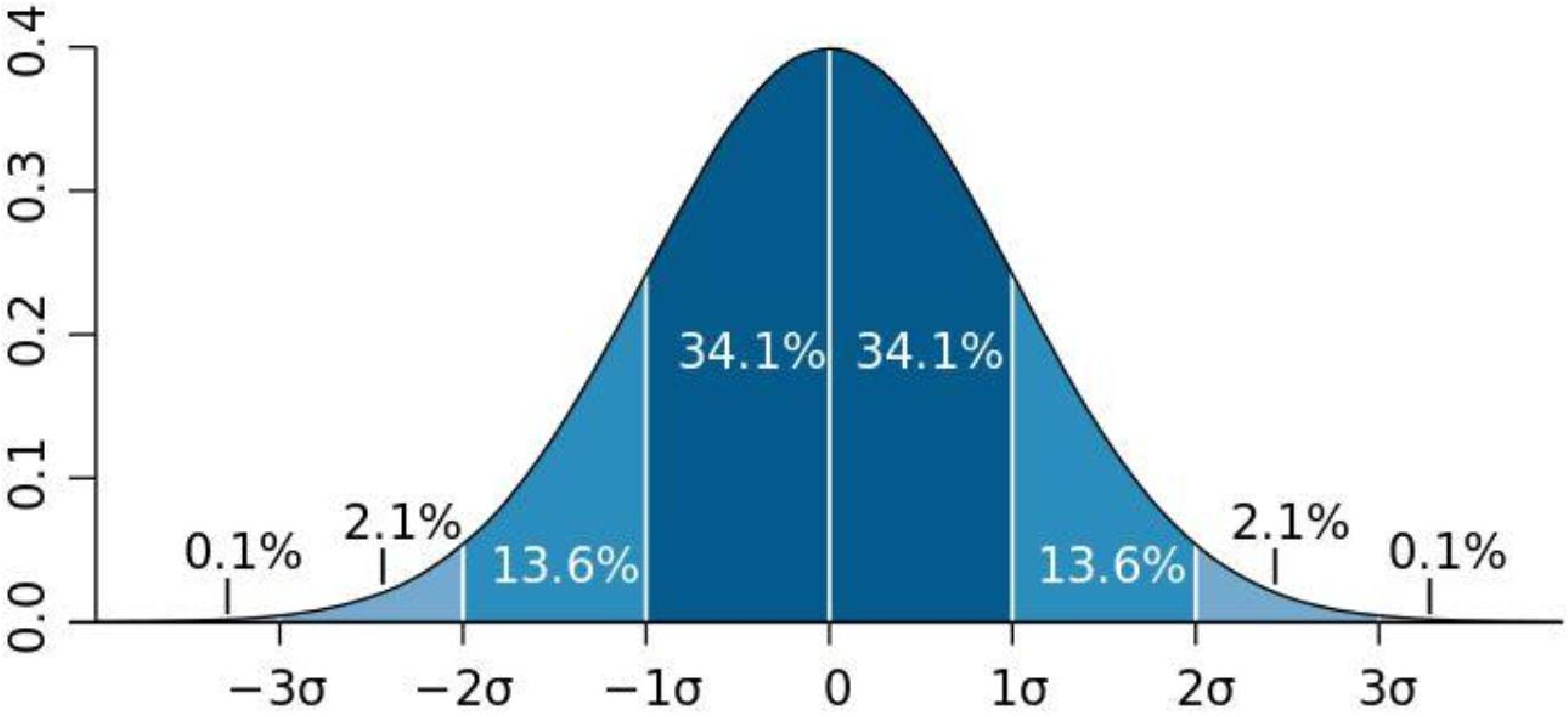
$$p(x \le a) = \Phi(a)$$

$$p(a < x \le b) = \Phi(b) - \Phi(a)$$

$$p(x > a) = 1 - \Phi(a)$$

$$p(x \le -a) = \Phi(-a) = 1 - \Phi(a)$$

$$p(|x| \le a) = 2\Phi(a) - 1$$


0.4
0.3
0.2
0.1
0.0
34.1%
34.1%
13.6%
13.6%
2.1%
2.1%
0.1%
0.1%
−3σ
−2σ
−1σ
0
1σ
2σ
3σ

# $x^2$分布（卡方分布）

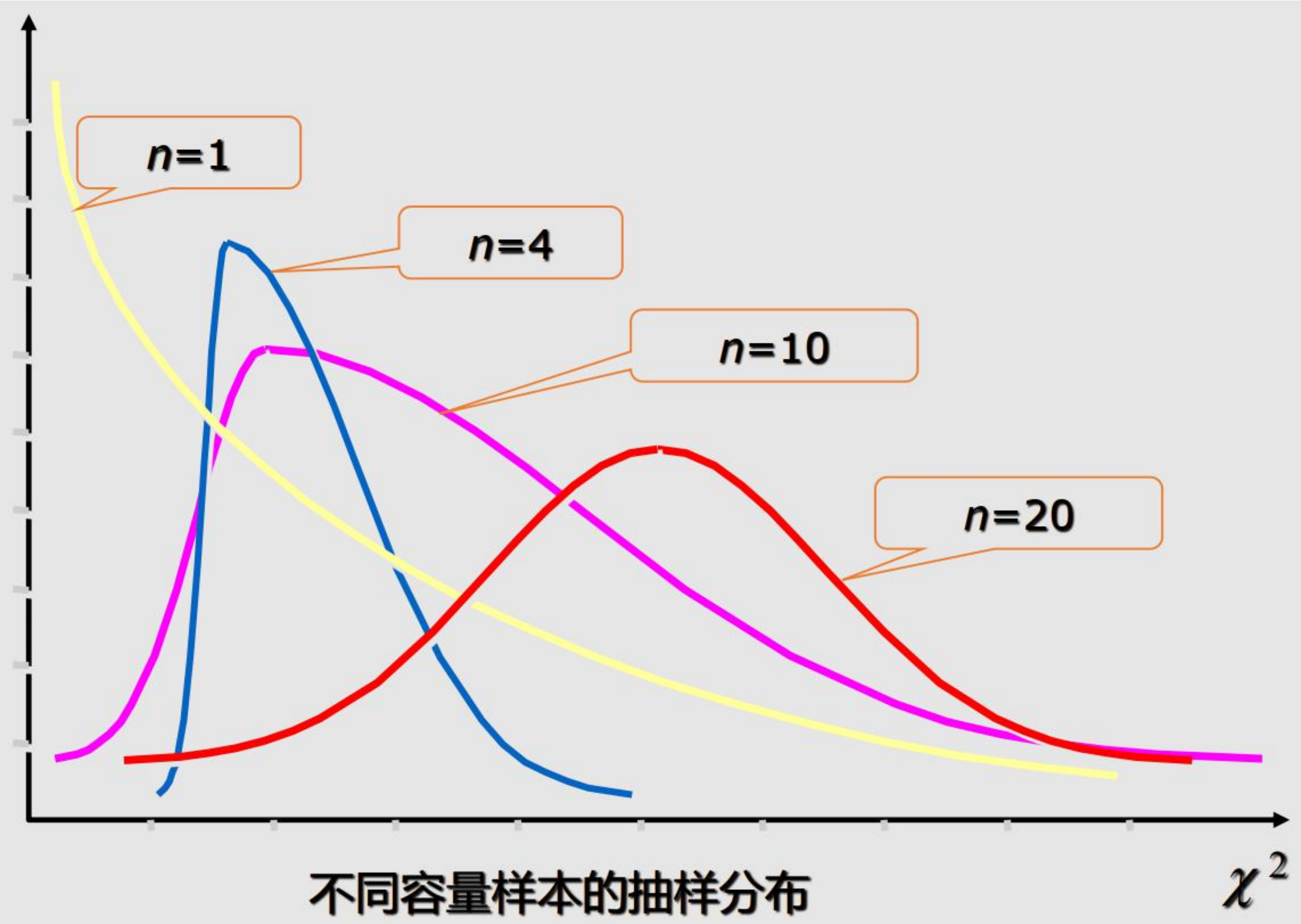
1. 由阿贝(Abbe) 于1863年首先给出，后来由海尔墨特(Hermert)和卡·皮尔逊(K·Pearson) 分别于1875年和1900年推导出来
2. 设 $X \sim N(\mu, \sigma^2)$ ，则 $z = \dfrac{X - \mu}{\sigma} \sim N(0,1)$

   令 $Y = z^2$ ，则 $Y$服从自由度为1的$\chi^2$分布，即

   $$Y \sim \chi^2(1)$$

3. 当总体 $X \sim N(\mu, \sigma^2)$ ，从中抽取容量为*n*的样本，则

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



n=1
n=4
n=10
n=20
$\chi^2$
不同容量样本的抽样分布

# $x^2$分布（卡方分布）

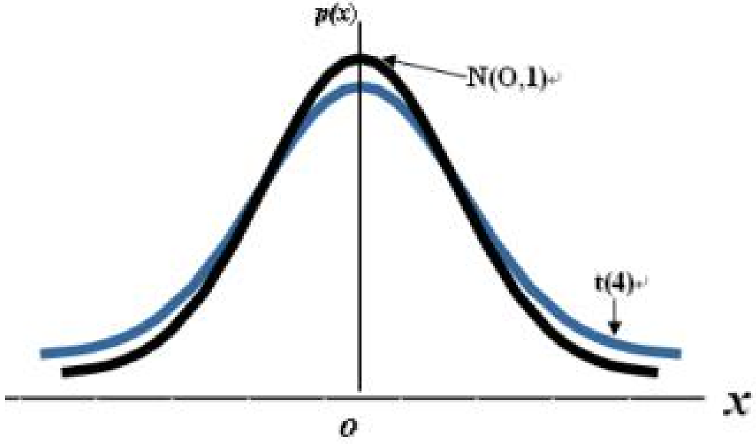
3、性质和特点

- 分布的变量值始终为正;
- 分布的形状取决于其自由度n 的大小，通常为不对称的正偏分布，但随着自由度的增大逐渐趋于对称;
- 常用于方差的估计和假设检验，以及列联分析中;
- 期望为：E($x^2$)=n ，方差为：D($x^2$)=2n (n为自由度）;
- 可加性：若U和V为两个独立的$x^2$分布随机变量，U ~ $x^2(n_1)$，V ~ $x^2(n_2)$,则U+V 这一随机变量服从自由度为$n_1 + n_2$的$x^2$分布;
- 当自由度增加到足够大时， $x^2$分布的概率密度曲线趋于对称，当n $\rightarrow \infty$时， $x^2$分布的极限分布是正态分布。

# t分布

1、定义：设随机变量$X \sim N(0,1)$，$Y \sim x^2(n)$，且X与Y独立，则$t=\frac{X}{\sqrt{\frac{Y}{n}}}$，其分布称为自由度为n的t 分布，记为t(n)。

2、图像：



# t分布

3、性质和特点:

- 当$n \geq 2$时，t分布的数学期望$E(t) = 0$，当$n \geq 3$时，t分布的方差$D(t) = \frac{n}{n-2}$。
- 自由度为1的t分布称为柯西分布。
- 随着自由度n的增加，t分布的密度函数愈来愈接近标准正态分布的密度函数。实际中，当$n \geq 30$时，t分布与标准正态分布就非常接近。
- t分布的提出对于统计学中小样本理论和应用有着重要的促进作用。

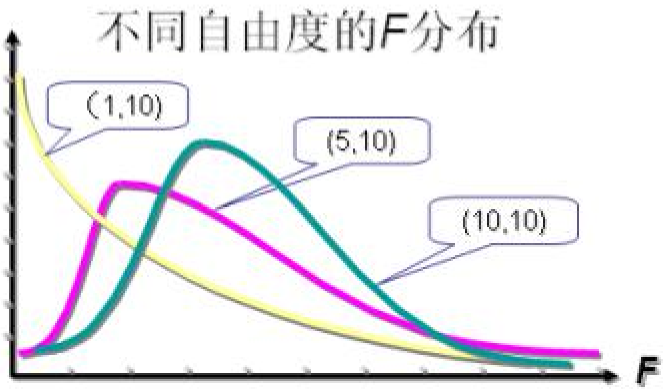
# F分布

1、定义：

若U为服从自由度为m的$x^2$分布，即$U \sim x^2(m)$，V 为服从自由度为n的$x^2$分布，即$V\sim x^2(n)$，且U和V 相互独立，

则$F=\frac{\frac{U}{m}}{\frac{V}{n}}$，称F 为服从自由度m和n的F 分布，记为$F\sim F(m,n)$，其中m为分子自由度，n为分母自由度。

2、图像：



# F分布

3、性质和特点:

- F分布的数学期望$E(t)=\frac{n}{n-2}$，$n>2$；方差 $D(t)=\frac{2n^2(m+n-2)}{m(n-2)(n-4)},n>4$
- F分布与t分布的关系：如果随机变量X服从$t(n)$分布，则$X^2$服从$F(1,n)$的F分布。
- F分布的应用广泛，如在方差分析、回归方程的显著性检验中都有重要地位。

# 相关分析

04

# 相关分析

问题起源：经济变量间的相互关系

- 确定性的函数关系 $Y = f(X)$
- 不确定性的统计关系—相关关系 $Y = f(X) + \varepsilon$ （ε为随机变量）
- 没有关系

【思考】如何进行相关分析？

从两个角度进行分析：

一、相关关系的描述
二、相关关系的度量

# 相关关系的描述——散点图

**相关关系的类型：**

1、从涉及的变量数量：

单相关

复相关（多重相关）

2、从变量相关关系的表现形式：

线性相关：散布图接近一条直线

非线性相关：散布图接近一条曲线

3、从变量相关关系变化的方向：

正相关：变量同方向变化，同增同减

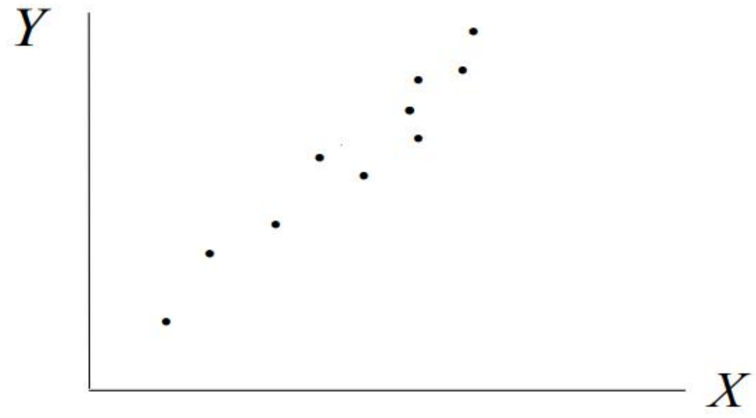
负相关：变量反方向变化，一增一减

4、从变量相关的程度看

完全相关

不相关

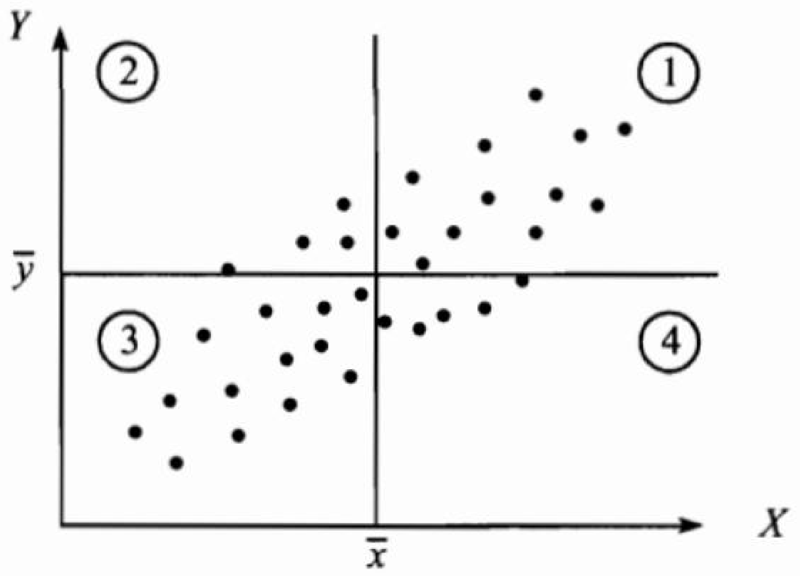
不完全相关



# 相关关系的度量——协方差

- Y 和 X 的协方差

$$\mathrm{Cov}(Y,X)=\frac{\sum_{i=1}^{n}(y_i-\bar{y})(x_i-\bar{x})}{n-1}$$




- Cov(Y,X) > 0　Y 和 X 正相关
- Cov(Y,X) < 0　Y 和 X 负相关
- 受度量单位的影响（不能反映变量间线性关系的强弱）

# 相关关系的度量——相关系数

1. Y和X 的相关系数

$$\mathrm{Cor}(Y,X)=\frac{\mathrm{Cov}(Y,X)}{s_y s_x}$$

$$=\frac{\sum(y_i-\bar{y})(x_i-\bar{x})}{\sqrt{\sum(y_i-\bar{y})^2\sum(x_i-\bar{x})^2}}$$

- Cov 协方差
- Cor 相关系数

# 相关关系的度量——相关系数

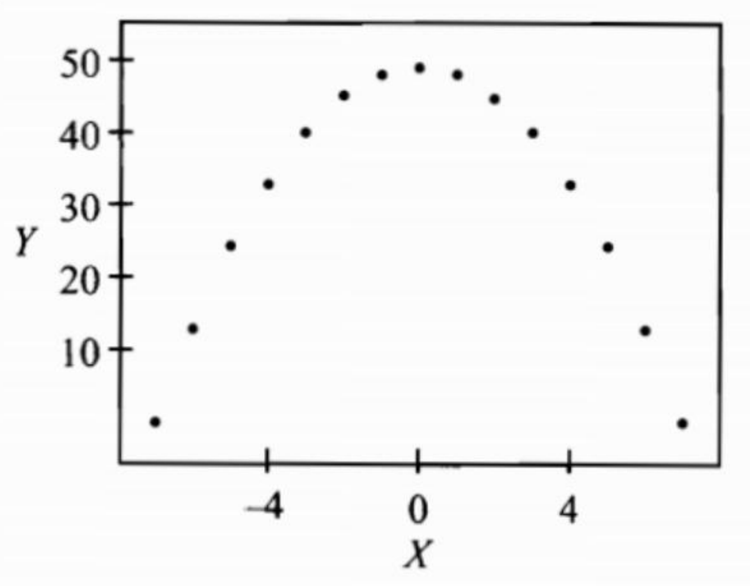
2、相关系数的特点：

- r 的取值范围是 [-1,1]
- $-1 \leq r < 0$，为负相关；$0 < r \leq 1$，为正相关
- $r = 0$，表明不存在线性相关关系
- $|r| = 1$，为完全相关；  $r = 1$，为完全正相关；  $r = -1$，为完全负相关
- $|r|$越趋于1表示关系越密切；$|r|$越趋于0表示关系越不密切

# 相关关系的度量——相关系数

- $\mathrm{Cor}(Y,X) = 0$
  - Y 和 X 之间没有线性相关性

| Y | X | Y | X |
|---|---|---|---|
| 1 | -7 | 49 | 1 |
| 14 | -6 | 46 | 2 |
| 25 | -5 | 41 | 3 |
| 34 | -4 | 34 | 4 |
| 41 | -3 | 25 | 5 |
| 46 | -2 | 14 | 6 |
| 49 | -1 | 1 | 7 |
| 50 | 0 | | |




- $Y = 50 - x^2$ $\quad and \quad Cor(Y, X) = 0$

# 相关关系的度量——相关系数

CDA数据分析师 CERTIFIED DATA ANALYST

- 相关系数易受到离群值的影响

| Y1 | X1 | Y2 | X2 | Y3 | X3 | Y4 | X4 |
|---|---|---|---|---|---|---|---|
| 8.04 | 10 | 9.14 | 10 | 7.46 | 10 | 6.58 | 8 |
| 6.95 | 8 | 8.14 | 8 | 6.77 | 8 | 5.76 | 8 |
| 7.58 | 13 | 8.74 | 13 | 12.74 | 13 | 7.71 | 8 |
| 8.81 | 9 | 8.77 | 9 | 7.11 | 9 | 8.84 | 8 |
| 8.33 | 11 | 9.26 | 11 | 7.81 | 11 | 8.47 | 8 |
| 9.96 | 14 | 8.1 | 14 | 8.84 | 14 | 7.04 | 8 |
| 7.24 | 6 | 6.13 | 6 | 6.08 | 6 | 5.25 | 8 |
| 4.26 | 4 | 3.1 | 4 | 5.39 | 4 | 12.5 | 19 |
| 10.84 | 12 | 9.13 | 12 | 8.15 | 12 | 5.56 | 8 |
| 4.82 | 7 | 7.26 | 7 | 6.42 | 7 | 7.91 | 8 |
| 5.68 | 5 | 4.74 | 5 | 5.73 | 5 | 6.89 | 8 |

- $Cor(Y_1, X_1) = Cor(Y_2, X_2) = Cor(Y_3, X_3) = Cor(Y_4, X_4) \approx 0.8$

# 相关关系的度量——相关系数

CDA数据分析师
CERTIFIED DATA ANALYST

| Y1 | X1 | Y2 | X2 | Y3 | X3 | Y4 | X4 |
|---|---|---|---|---|---|---|---|
| 8.04 | 10 | 9.14 | 10 | 7.46 | 10 | 6.58 | 8 |
| 6.95 | 8 | 8.14 | 8 | 6.[illegible] | [illegible] | [illegible] | [illegible] |
| 7.58 | 13 | 8.74 | 13 | 12[illegible] | [illegible] | [illegible] | [illegible] |
| 8.81 | 9 | 8.77 | 9 | 7[illegible] | [illegible] | [illegible] | [illegible] |
| 8.33 | 11 | 9.26 | 11 | 7[illegible] | [illegible] | [illegible] | [illegible] |
| 9.96 | 14 | 8.1 | 14 | 8[illegible] | [illegible] | [illegible] | [illegible] |
| 7.24 | 6 | 6.13 | 6 | 6[illegible] | [illegible] | [illegible] | [illegible] |
| 4.26 | 4 | 3.1 | 4 | 5[illegible] | [illegible] | [illegible] | [illegible] |
| 10.84 | 12 | 9.13 | 12 | 8[illegible] | [illegible] | [illegible] | [illegible] |
| 4.82 | 7 | 7.26 | 7 | 6[illegible] | [illegible] | [illegible] | [illegible] |
| 5.68 | 5 | 4.74 | 5 | 5[illegible] | [illegible] | [illegible] | [illegible] |