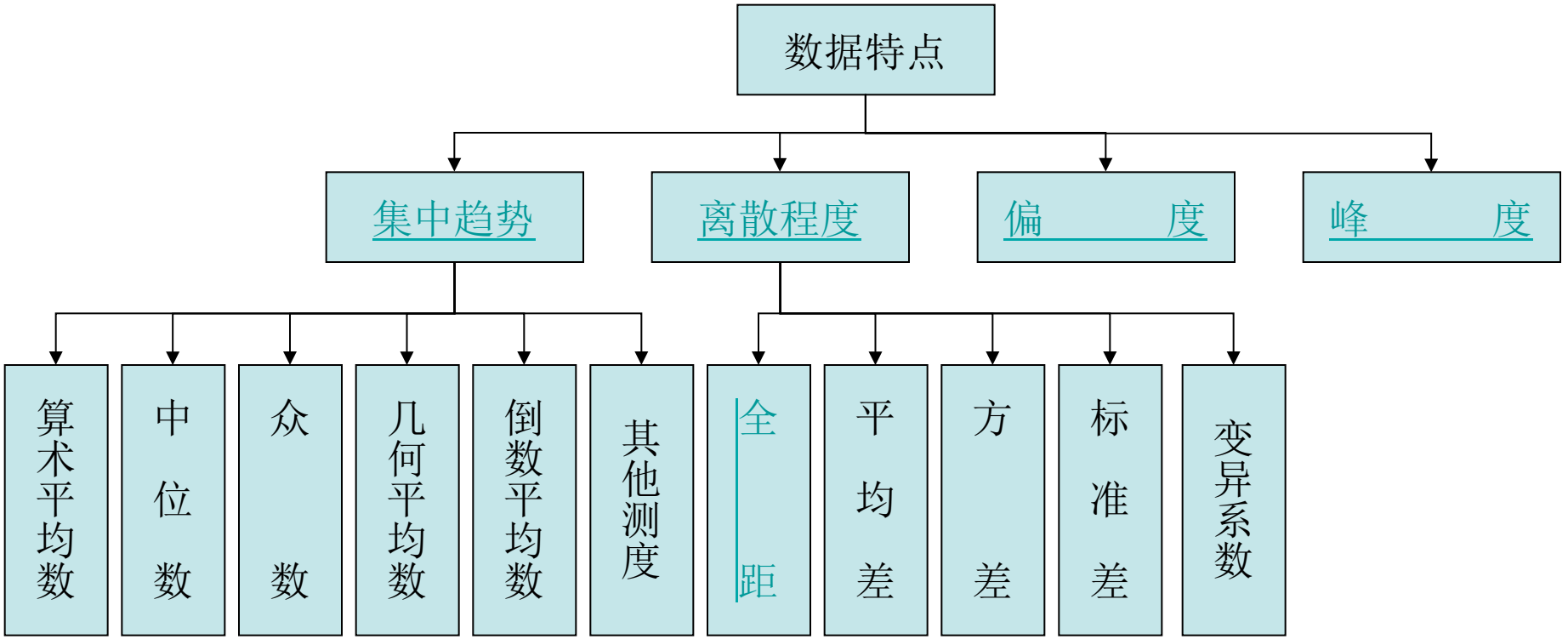
第二章 常用统计指标

引言：数据分布的特征及特征量数



一、集中量数——算术平均数（arithmetic mean）

1、定义

算术平均数是观察值的总和除以观测值的个数所得的商数。统计学上也称为原点矩。

2、总体平均数与样本平均数

$$①\mu=\frac{\sum X}{N}$$

$$②\bar{X}=\frac{\Sigma X}{n}=\frac{X_1+X_2+\mathrm{L}\ +X_n}{n}$$

例：某校初三年级5个学生的一次数学统计成绩分别是：46、54、42、46、32，问这5个学生的数学成绩平均是多少？

解：$\bar{X}=\frac{\Sigma X}{n}$

$$=\frac{46+54+42+46+32}{5}=44$$

一、集中量数——算术平均数（arithmetic mean）

3、加权算术平均数

当各个数据出现的次数不等时，计算平均数时，应该用加权算术平均数。

①计算公式

$$\bar{X} = \frac{\Sigma X_i W_i}{\Sigma W_i} = \frac{X_1 W_1 + X_2 W_2 + X_3 W_3 + \mathrm{L} + X_k W_k}{W_1 + W_2 + W_3 + \mathrm{L} + W_k}$$

②未分组数据

③分组数据

一、集中量数——算术平均数（arithmetic mean）

4、平均数的特点

①在一组数据中，每一个数据加上一个常数C，则所得的平均数为原来的平均数加上常数C。即：

$$\frac{\sum_{i=1}^{n}(x_i+C)}{n}=\frac{\sum_{i=1}^{n}x_i}{n}+\frac{\sum_{i=1}^{n}C}{n}=\bar{X}+C$$

②在一组数据中，每一个数据乘上一个常数C，则所得的平均数为原来的平均数乘上常数C。即：

$$\frac{\sum_{i=1}^{n}(x_iC)}{n}=\frac{C\sum_{i=1}^{n}x_i}{n}=C\bar{X}$$

③离均差的和等于0。即：$\sum(x_i-\bar{X})=0$

①观测值与任意常数C的离差平方和，不小于离均差平方和。即：$\sum(x_i-C)^2\geqslant\sum(x_i-\bar{X})^2$

证明如下：

$$x_i-C=(x_i-\bar{X})-(\bar{X}-C)$$

$$\sum(x_i-C)^2=\sum(x_i-\bar{X})^2-2(\bar{X}-C)\sum(x_i-\bar{X})+\sum(\bar{X}-C)^2$$

$$=\sum(x_i-\bar{X})^2+n(\bar{X}-C)^2\geqslant\sum(x_i-\bar{X})^2$$

一、集中量数——算术平均数（arithmetic mean）

5、平均数的优点

①反应灵敏　　②计算严密、简单、易理解
③受抽样变动的影响小　　④适于进一步用代数方法演算

6、平均数的缺点

①易受极端数据的影响。此时，可考虑用中位数或修剪平均数（trimmed mean）。
②若出现模糊不清的数据时，无法计算平均数。此时，一般采用中位数描述其集中趋势。

7、计算和使用平均数的原则

①同质性原则。
②平均数与个体数值相结合的原则。
③平均数与标准差、方差相结合的原则。

二、集中量数——中位数（median）

1、定义

中位数是将一批数据从小至大排列后位次居中的数据值，符号为**Mdn**，反映一批观察值在位次上的平均水平。

2、适用条件：

适合各种类型的资料。尤其适合于①大样本偏态分布的资料； ②资料有不确定数值；③资料分布不明等。

3、计算方法：

①原始数据求中数

先将观察值按从小到大顺序排列，再按以下公式计算中位数的位置：

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

②分组数据求中数

三、集中量数————众数（mode）

1、定义

在一组数据中出现次数（或频数）最多的观察值。记为**Mo**。

2、原始数据求众数

例：求8、9、9、9、11、12、12、12、14的众数。

解：Mo＝9和12

3、分组数据求众数

①以最高次数组的组中值作为众数。

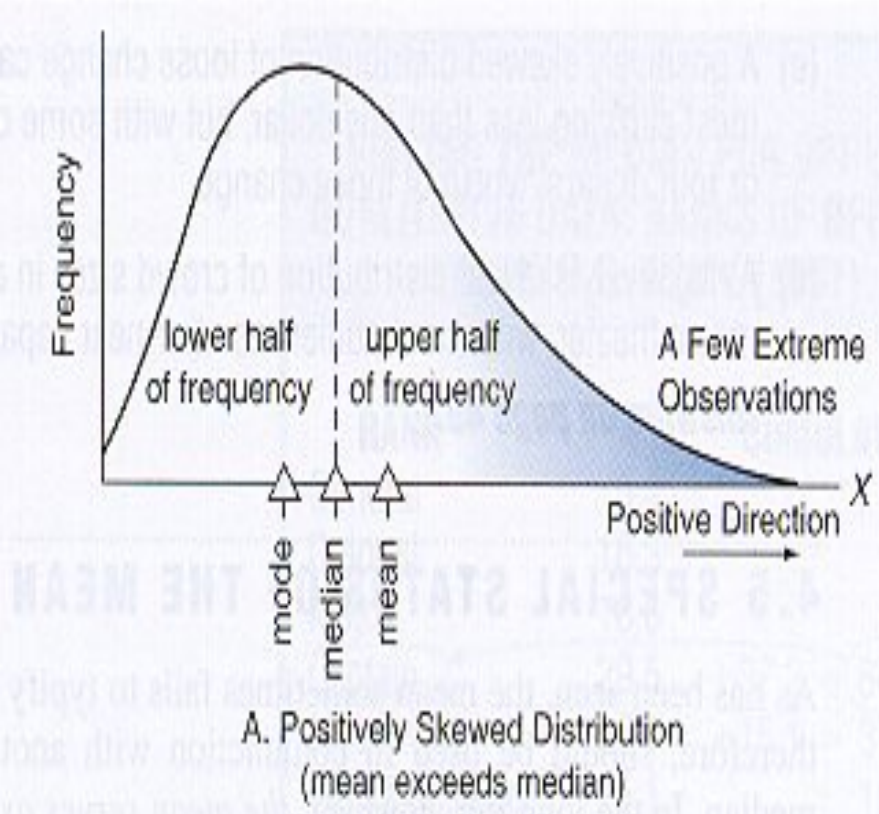
②皮尔逊经验公式：$Mo = 3Mdn - 2\bar{X}$

例：以求中位数的分组为例。

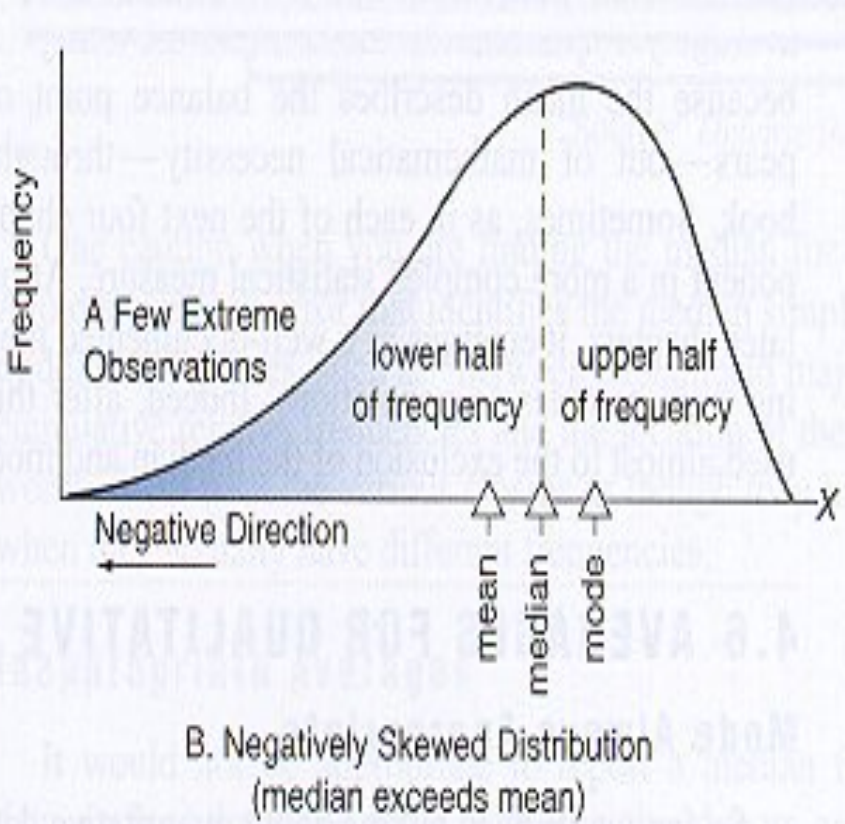
解：Mo＝3×84.875-2 ×85.6585＝83.308

四、均数、中位数、众数三者关系

正态分布时：　均数＝中位数＝众数
正偏态分布时：均数>中位数>众数
负偏态分布时：均数<中位数<众数



A. Positively Skewed Distribution
(mean exceeds median)



B. Negatively Skewed Distribution
(median exceeds mean)

五、集中量数——几何平均数（geometric mean）

1、定义：

n个数据乘积的n次方根，称为几何平均数。记为：$\bar{X}_G$

$$\bar{X}_G = \sqrt[n]{X_1 X_2 \mathrm{L}\ X_n}$$

2、应用情形：

①数据按一定的比例关系变化，如求平均增长率或对心理物理学中的等距与等比量表实验的数据处理均应用几何平均数。

②当一组数据中存在极端数据，分布呈偏态时，算术平均数不能很好地反映数据的典型情况，此时应使用几何平均数或其他集中量数（中数、众数等）来反映数据的典型情况。

3、计算例子。

六、集中量数——调和平均数（harmonic mean）

1、定义：

将各数据取倒数求平均，然后再取倒数，所得值即为调和平均数，又称为倒数平均数。记为：M_H

$$M_H = \frac{1}{\frac{1}{N}\left(\frac{1}{X_1}+\frac{1}{X_2}+\cdots+\frac{1}{X_n}\right)} = \frac{N}{\sum\frac{1}{X_i}}$$

2、适用情境：求平均学习速度

①学习任务的工作量相同，所用时间不等，求平均学习速度

②学习任务的时间相同而工作量不等，求平均学习速度

七、集中趋势的其他测度指标

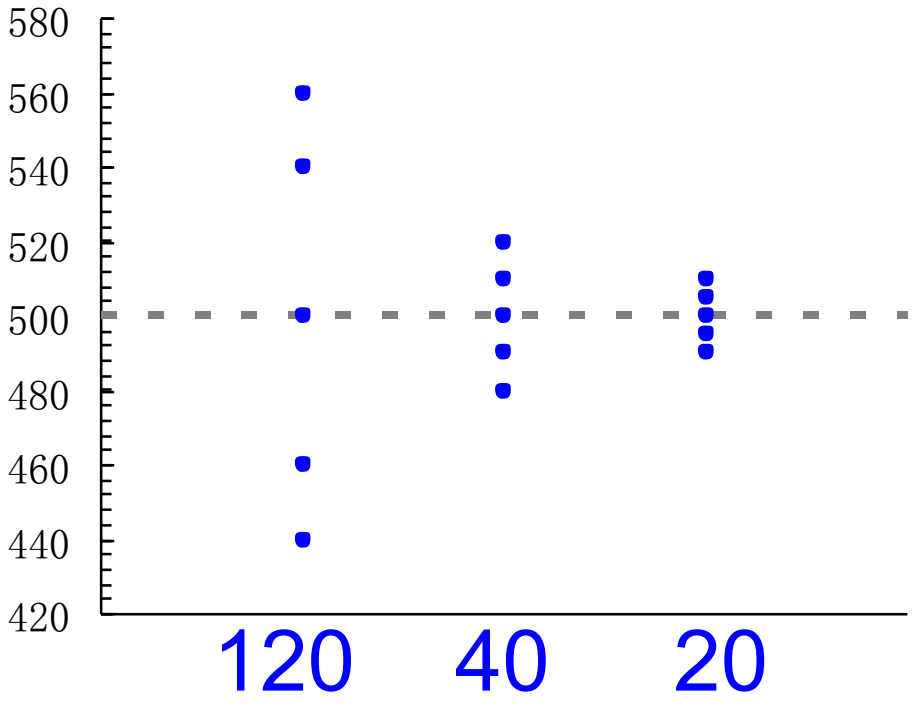
- Huber的M估计量
- Hampel的M估计量
- Tukey的M估计量
- Andrew的M估计量

八、差异量数——全距(Range)

$$R = X_{\max} - X_{\min}$$

优点：简便

缺点：1. 只利用了两个
极端值
2.*n*大，R也
3.不稳定



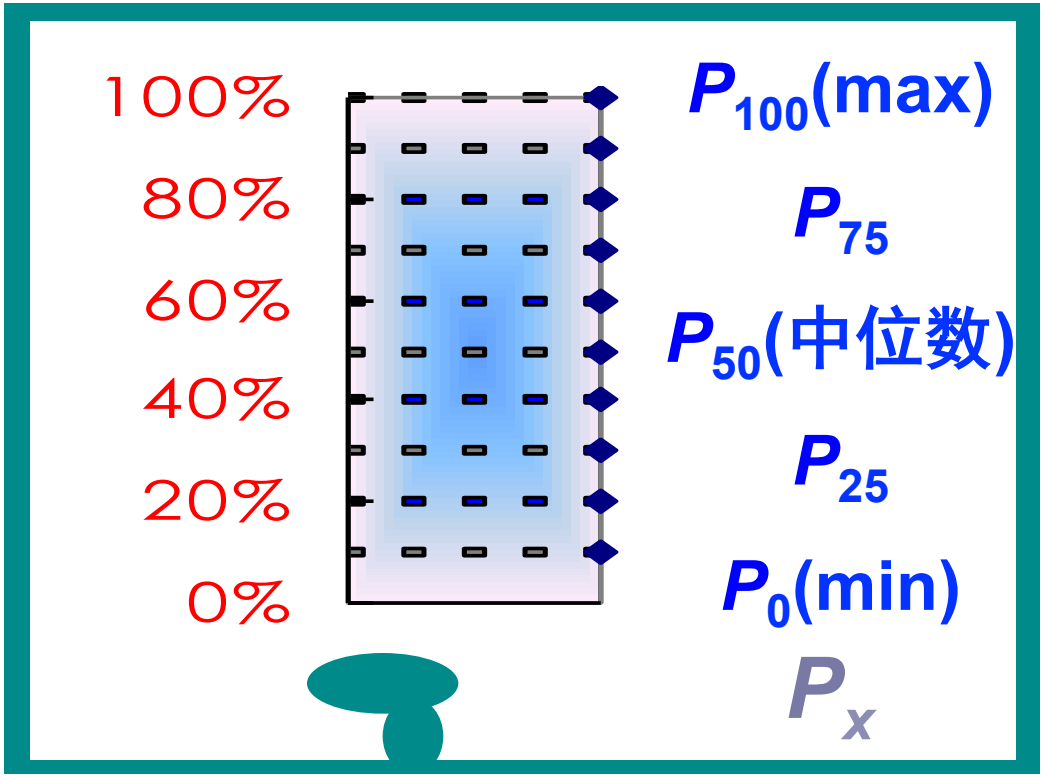
九、差异量数——百分位差与四分位差

百分位数 ：数据从小到大排列;在百分尺度下，所占百分比对应的值。记为P_x

百分位差：P_{90}-P_{10}

四分位间距：
$IQR = P_{75} - P_{25}$

四分位差（quartile deviation）：$QD = IQR/2$



十、差异量数——平均差（average deviation）

1、定义：

平均差 （**average deviation或mean deviation**）是次数分布中所有原始数据与平均数绝对离差的平均值。记为*A.D.或M.D.*。

2、公式： $A.D. = \dfrac{\sum |x_i - \bar{X}|}{n}$

3、计算

①未分组数据

②分组数据

十一、差异量数——方差（variance）

1、定义

方差（variance）也称**均方差**（mean square deviation），样本观察值的离均差平方和的均值。表示一组数据的平均离散情况。

2、公式

离均差和 $\sum(X-\mu)=0$

离均差平方和(sum of square) $SS=l_{xx}=\sum(X-\mu)^2$

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

样本方差 $S^2=\dfrac{\sum(X-\bar{X})^2}{n-1}=\dfrac{\sum X^2-\left(\sum X\right)^2/n}{n-1}$

样本方差为什么要除以（$n-1$）

与自由度（degrees of freedom）有关。

自由度是数学名词，在统计学中，n个数据如不受任何条件的限制，则n个数据可取任意值，称为有n个自由度。若受到k个条件的限制，就只有（$n-k$）个自由度了。计算样本方差时， n个变量值本身有n个自由度。但受到样本均数的限制，任何一个“离均差”均可以用另外的（$n-1$）个“离均差”表示，所以只有（$n-1$）个独立的“离均差”。因此只有（$n-1$）个自由度。

十二、差异量数——标准差（standard deviation）

1、定义

标准差（standard deviation）即方差的正平方根；其单位与原变量X的单位相同。

2、公式

$$\text{总体标准差}\sigma=\sqrt{\frac{\sum(x_i-\mu)^2}{N}}$$

$$\text{样本标准差}S=\sqrt{\frac{\sum(x_i-\bar{X})^2}{n-1}}$$

$$\text{频数表样本标准差}S=\sqrt{\frac{\sum(x_i-\bar{X})^2\cdot f}{n-1}}$$

3、计算

①未分组数据

②分组数据

十二、差异量数——标准差（standard deviation）

4、标准差的性质

①每一个观测值都加上一个相同常数C后，计算得到的标准差等于原标准差。即

$$y_i = x_i + c, \text{则} S_y = S_x$$

②每一个观测值都乘以一个相同常数C（C≠0），则所得的标准差等于原标准差乘以这个常数。即

$$y_i = c \cdot x_i, \text{则} S_y = c \cdot S_x$$

③每一个观测值都乘以一个相同常数**C（C≠0）**，再加上一个常数d后，则所得的标准差等于原标准差乘以这个常数C。即

$$y_i = c \cdot x_i + d, \text{则} S_y = c \cdot S_x$$

例3.3 未分组数据求方差和标准差：

X_i	$X_i-\overline{X}=x$	x^2	X_i^2
6	0	0	36
5	-1	1	25
7	1	1	49
4	-2	4	16
6	0	0	36
8	2	4	64
N=6 $\sum X_i=36$	$\sum x=0$	$\sum x^2=10$	$\sum X_i^2=226$

$\sigma^2=10/6=1.67$ ， $\sigma=1.29$

用原始数据直接求方差和标准差:

$$\sigma^2 = \frac{\sum X_i^2}{N} - \left(\frac{\sum X_i}{N}\right)^2 = \frac{N\sum X_i^2 - \left(\sum X_i\right)^2}{N^2},$$

$$\sigma = \sqrt{\frac{\sum X_i^2}{N} - \left(\frac{\sum X_i}{N}\right)^2}$$

上例中

$$\sigma^2 = \frac{226}{6} - \left(\frac{36}{6}\right)^2 = 1.67,$$

$$\sigma = \sqrt{1.67} = 1.29$$

课堂练习

1、分别求下列各组数据的方差、标准差

（1）15，16，13，11，12，10，11

（2）5，6，3，1，2，0，1

（3）10，12，6，2，4，0，2

方差与标准差的意义

方差与标准差是表示一组数据离散程度的最好的指标,是高效差异量。

（1）反应灵敏。

（2）由计算公式严格确定；

（3）容易计算；

（4）适合代数运算；

（5）受抽样变动的影响小，即不同样本的标准差或方差比较稳定；

（6）简单明了；

（7）具有可加性。可以把总变异分解为不同来源的变异。

（8）各变量值对均值的方差小于对任意数的方差。即：

$$\sigma^2 < D^2$$

十三、差异量数——差异系数(coefficient of variation)

1、含义与公式：

差异系数又叫变异系数，它是一种相对差异量。记为：CV。

$$CV = \frac{S}{\overline{X}} \times 100\%$$

2、适用条件：

①观察指标单位不同，如身高、体重

②同单位资料，但均数相差悬殊

3、计算实例

十四、差异系数指标小结

1．全距较粗，适合于任何分布

2．标准差与均数的单位相同，最常用，适合于近似正态分布

3．变异系数主要用于单位不同或均数相差悬殊资料

集中量数和差异量数分别反映资料的不同特征，常配套使用：

如 正态分布：均数、标准差；

偏态分布：中位数、四分位半

十五、偏态量数、峰度量数

偏态系数是对分布偏斜程度的测度，其计算公式为：

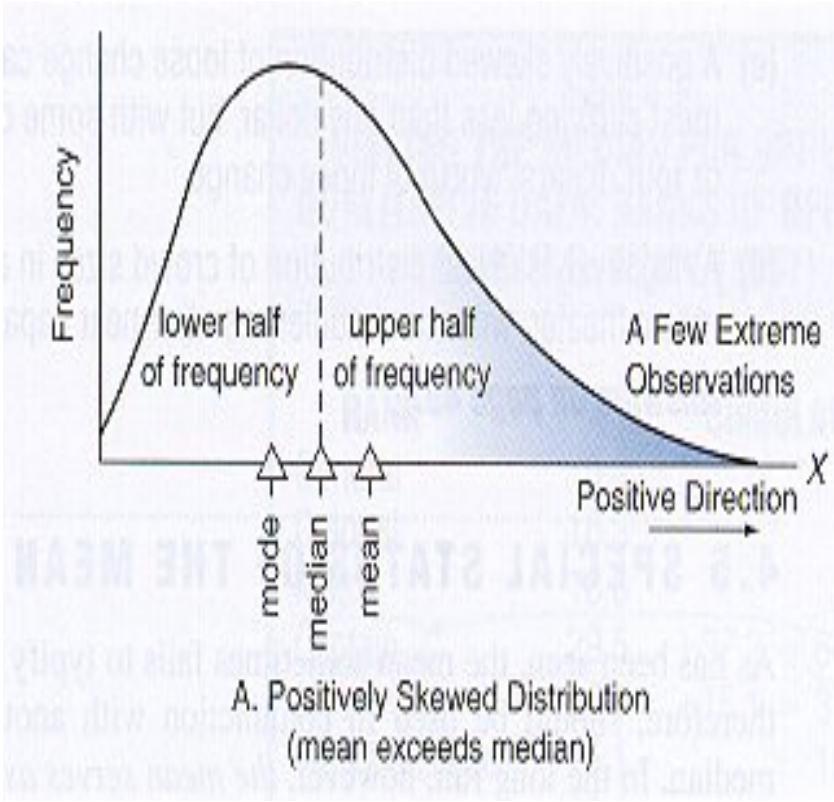
$$a_3 = \frac{\sum_{i=1}^{K}\left(X_i - \overline{X}\right)^3}{N\sigma^3}$$

峰度系数是对分布集中趋势高峰形状的测度，其计算公式为：

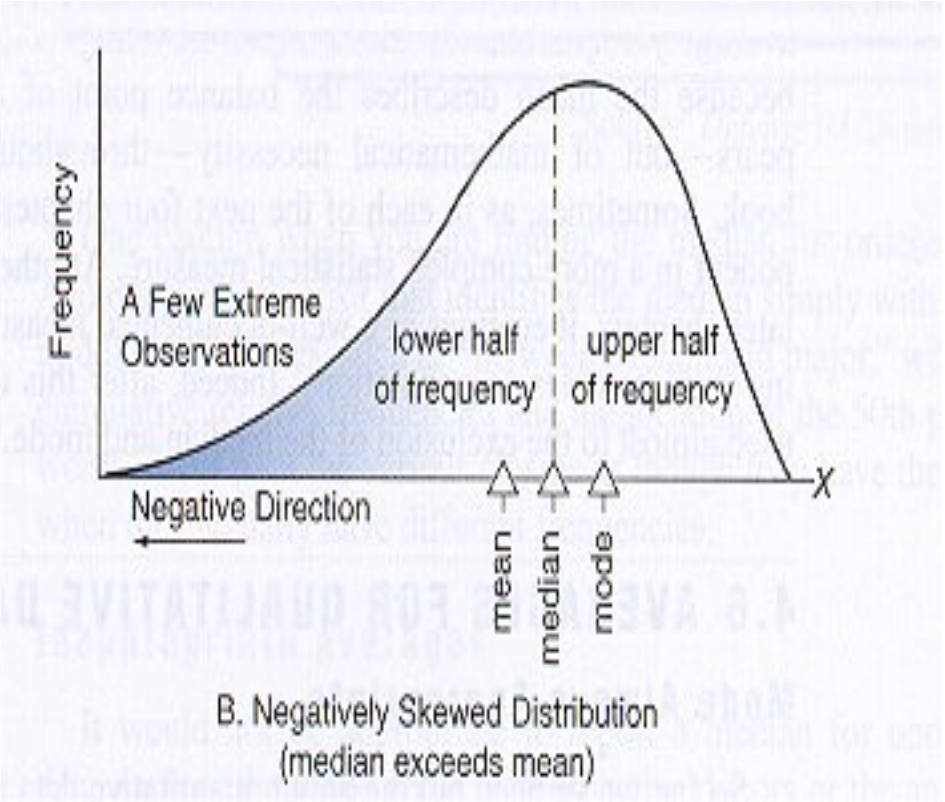
$$a_4 = \frac{\sum_{i=1}^{K}\left(X_i - \overline{X}\right)^4}{N\sigma^4} - 3$$

偏度的含义

- 偏度（**Skewness**）是对分布偏斜方向和程度的测度。



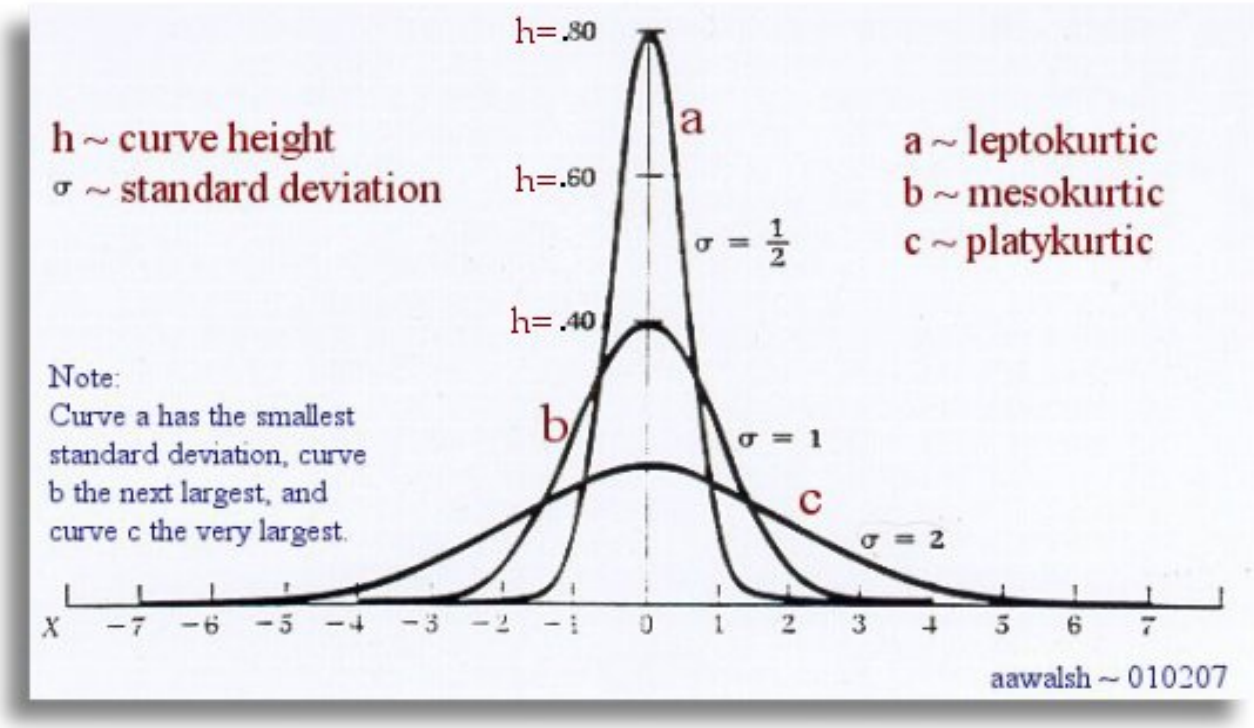
A. Positively Skewed Distribution
(mean exceeds median)



B. Negatively Skewed Distribution
(median exceeds mean)

峰度的含义

- 峰度（Kurtosis）是分布集中趋势高峰的形状。它通常是与正态分布相比较而言。
  - 若分布的形态比正态分布更瘦更高，则称为尖峰分布。若比正态分布更矮更胖，则称为平峰分布。



返回

标准分数(Standard score) , Z分数.

$$Z_i = \frac{X_i - \mu}{\sigma}$$

$$Z_i = \frac{x_i - \bar{x}}{S}$$

标准分数可以给出各数值在一组数据中的相对位置。

例某班平均成绩为90分，标准差为3分，甲生得94.2分,乙生得89.1分,求甲乙二学生的Z分数各是多少?

解:$Z_{甲}=(94.2-90)/3=1.4$

$Z_{乙}=(89.1-90)/3=-0.3$

标准分数的平均值为0,标准差为1。

Z分数的应用:

(1)比较分属性质不同的观测值在各自数据分布中相对位置的高低。

如:某人 $Z_{身高1.70}=0.5$, $Z_{体重65}=1.2$, 则该人在某团体中身高稍偏高,而体重更偏重些。

(2) 当已知各不同质的观测值的次数分布为正态时,可用Z分数求不同的观测值的总和或平均值,以表明在总体中的位置。

表3.1 利用Z分数求总和

科目	原始分数		全体考生		Z 分数	
	甲	乙	平均数	标准差	甲	乙
语文	85	89	70	10	1.500	1.900
政治	70	62	65	5	1.0	-0.600
外语	68	72	69	8	-0.125	0.375
数学	53	40	50	6	0.500	-1.670
理化	72	87	75	8	-0.375	1.500
总计	348	350			2.500	1.505

(3)表示标准测验分数

经过标准化的测验,如果其常模分数分布接近正态分布,常常要转换成正态标准分数。

$$Z'=aZ+b$$

Z’为正态标准分数,$Z=(X-\bar{X})/\sigma$, a,b为常数, σ为测验常模的标准差。

如：（WAIS）韦氏常人智力量表：IQ=15Z+100；

比奈--西蒙智力测验：Z’=16Z+100;

普通分类测验(AGCT) Z’=20Z+100

(4)异常值(极端值)的取舍

一个正态分布中,平均数上下一定的标准差处,包含有确定百分数的数据个数。如上下三个标准差内包含99.73%的数据个数。所以，如果有一个数据的取值落在平均数加减三个标准差之外，则在整理数据时，可将此数据作为异常值加以舍弃。

切贝谢夫定理：

在任意一个数据集中，至少有（1-1/Z²）的数据项与平均数的距离在特定数目个标准差之内，其中z是任意大于1的值。

Z=2,3,4个标准差时,根据切贝谢夫定理:

- 至少75%的数据项与平均数的距离在Z=2个标准差之内;
- 至少89%的数据项与平均数的距离在Z=3个标准差之内;
- 至少94%的数据项与平均数的距离在Z=4个标准差之内。

在正态分布的情况下标准差与平均数之间有一定关系：

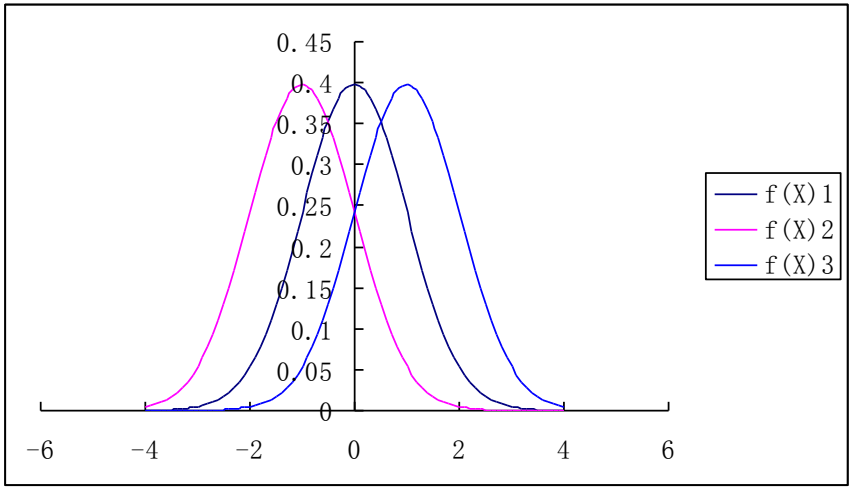
$\overline{X} \pm 1\sigma$ 或 $\bar{x} \pm 1S$ 包含总数目的68.26%

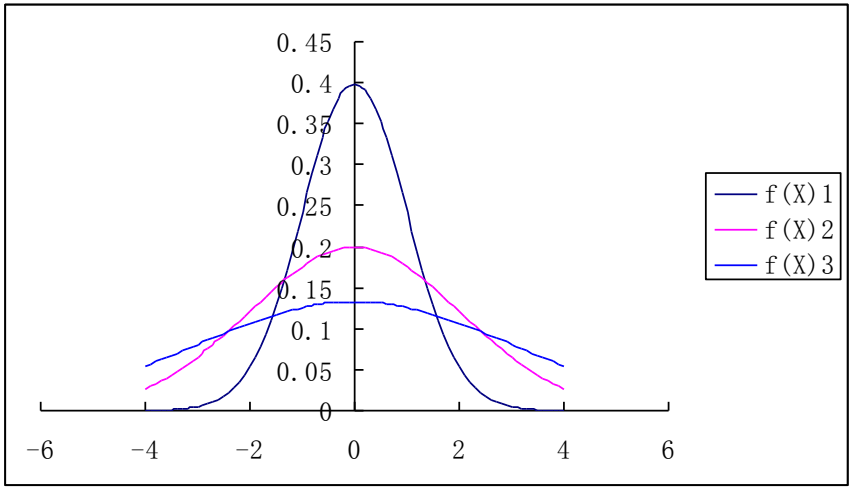
$\overline{X} \pm 1.96\sigma$ 或 $\bar{x} \pm 1.96S$ 包含总数目的95%

$\overline{X} \pm 2.58\sigma$ 或 $\bar{x} \pm 2.58S$ 包含总数目的99%

集中趋势、离中趋势含义

- 集中趋势（ *central tendency* ）与离中趋势是次数分布的两个基本特征。
- 集中趋势就是指数据分布中大量数据向某方向集中。
  - 描述数据集中趋势的特征量数称为集中量数。
- 离中趋势是指数据分布中数据彼此分散的程度。也称为离散程度。
  - 描述数据集中趋势的特征量数称为差异量数。





返回

未分组数据求加权算术平均数的例子

例：某校五年级有八个平行班，一次语文统考成绩情况如下表，请计算该校五年级这次语文统考的总平均分。

班级	1	2	3	4	5	6	7	8
人数	53	55	48	38	35	50	54	65
平均分	75	77	72	81	83	74	71	69

解：$\bar{X} = \dfrac{\sum X_i W_i}{\sum W_i}$

$$= \frac{75\times 53 + 77\times 55 + \mathrm{L} \ + 69\times 65}{53+55+48+\mathrm{L} \ +65}$$

$$= 74.5427$$

返回

分组数据求算术平均数的例子

例：求下列次数分布表的算术平均数。

组　别	组中值	次数
65-69	67	1
60-64	62	4
55-59	57	6
50-54	52	8
45-49	47	16
40-44	42	24
35-39	37	34
30-34	32	21
25-29	27	16
20-24	22	11
15-19	17	9
10-14	12	7
合　计		157

解：$\bar{X} = \frac{\Sigma X_i W_i}{\Sigma W_i} = \frac{\Sigma X_i f_i}{n}$

$$= \frac{67\times 1 + 62\times 4 + 57\times 6 + \mathrm{L}\ \ + 12\times 7}{157}$$

$$= 36.14$$

返回

未分组数据求中位数的例子

1、数据的个数为奇数，中数附近无重复。

例：求5、2、9、8、13的中数。

解：2、5、8、9、13

Mdn=8

2、数据的个数为偶数，中数附近无重复。

例：求11、2、9、8、12、20的中数。

解：2、8、9、11、12、20

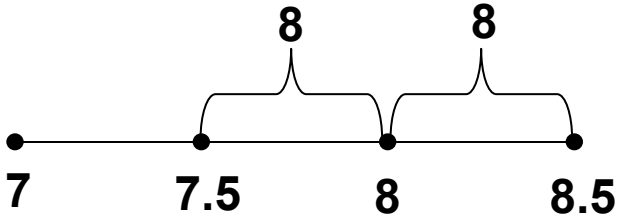
Mdn=（9+11）/2＝10

3、中数附近有重复。

例：求5、2、8、8、13的中数。

解：2、5、8、8、13

Mdn=（7.5+8）/2＝7.75



返回

分组数据求中位数的例子

例：求下列次数分布表的中位数。

组　别	次数	累积次数
93-95	5	41
90-92	4	36
87-89	5	32
84-86	12	27
81-83	11	15
78-80	4	4
合　计	41	

计算公式：$\text{Mdn}=L_b+\dfrac{\frac{N}{2}-F_b}{f}\cdot i$

式中：L_b为中数所在组的精确下限

f为中数所在组的次数

F_b为中数所在组的下一组的累积次数

N为总次数；i为组距

确定中数所在组：

①计算$\dfrac{N}{2}$

②将$\dfrac{N}{2}$与累积次数由下向上进行比较

③第一个大于或等于$\dfrac{N}{2}$的累积次数所在组即为中位数所在组。

解：$\text{Mdn}=83.5+\dfrac{20.5-15}{12}\times 3$

$=84.875$

返回

求几何平均数的例子

例1：有一研究者想研究介于S1与S2二感觉之间的感觉的物理刺激是多少，随机抽10名被试，让其调节一个可变的物理量的刺激，使其产生的感觉恰介于S1与S2之间，然后测量所调节刺激的物理量。10名被试的结果如下：

5.7、6.2、6.7、6.9、7.5、8.0、7.6、10.0、15.6、18.0

解：$X_G = \sqrt[10]{5.7 \times 6.2 \times \mathrm{L} \times 18.0} = 8.55$

例2：下面是某校几年来毕业生的人数，问平均增加率是多少？

年份	1981	1982	1983	1984	1985	1986
人数	760	810	930	1050	1120	1230

解：$\bar{X}_G = \sqrt[5]{\frac{810}{760} \times \frac{930}{810} \times \mathrm{L} \times \frac{1230}{1120}} = 1.1011$

$1.1011 - 1 = 0.1011 = 10.11\%$

返回

工作量相同，时间不等，求平均学习速度

例：有一学生15分钟学会生词30个，后10分钟学会生词也是30个。问该生每分钟平均学会多少？（或平均学习速度是多少？）

学会生字数（个）	所用时间（分钟）	平均每分钟学会字数
30	**15**	**30**÷15＝2
30	**10**	**30**÷10＝3

$$\text{解：}\ M_H = \frac{1}{\frac{1}{2}\left(\frac{1}{2}+\frac{1}{3}\right)} = \frac{2}{\frac{5}{6}} = 2.4\text{字/分钟}$$

返回

时间相同而工作量不等，求平均学习速度

例：一个学习实验结果如下表所示，求平均解题速度。

被试	解题数	所用时间（小时）	单位时间工作量
1	24	2	12
2	20	2	10
3	16	2	8
4	12	2	6
5	8	2	4
6	4	2	2

解：$M_H = \dfrac{1}{\frac{1}{6}(\frac{1}{12}+\frac{1}{10}+\frac{1}{8}+\frac{1}{6}+\frac{1}{4}+\frac{1}{2})} = 4.9$题/小时

返回

未分组数据计算平均差

例：有5名被试的错觉实验结果如下，求其平均差。

被　试	1	2	3	4	5
错觉量（要ms）	16	18	20	22	17

解：$A.D. = \dfrac{\sum |x_i - \bar{X}|}{n}$

$$= \frac{|16-18.6| + |18-18.6| + |20-18.6| + |22-18.6| + |17-18.6|}{5}$$

$$= 1.92$$

返回

分组数据计算平均差

例：一次划线实验误差分布如下（单位mm），求其平均差。

组别	组中值	f
55-59	57	1
50-54	52	3
45-49	47	5
40-44	42	8
35-39	37	16
30-34	32	15
25-29	27	9
20-24	22	6
15-19	17	3
10-14	12	2
∑		68

解：$A.D. = \dfrac{\sum |x_i - \bar{X}| \cdot f}{n}$

$$= \frac{|57-33.91|\times 1 + |52-33.91|\times 3 + \mathrm{L} + |12-33.91|\times 2}{68}$$

$$= 7.56$$

返回

未分组数据计算方差、标准差

例：6名被试的一项测试成绩如下，试求其方差与标准差。

被试	成绩
1	6
2	5
3	7
4	4
5	6
6	8

解：$S=1.4142$

$S^2=2$

$\sigma=1.2910$

$\sigma^2=1.6667$

返回

分组数据计算方差、标准差

例：一次划线实验误差分布如下（单位mm），求其方差、标准差。

组别	组中值	f
96-98	97	2
93-95	94	3
90-92	91	4
87-89	88	8
84-86	85	11
81-83	82	17
78-80	79	19
75-77	76	14
72-74	73	10
69-71	70	7
66-68	67	3
63-65	64	1
60-62	61	1
∑		100

解：$S=7.1488$

$S^2=51.1055$

$\sigma=7.1130$

$\sigma^2=50.5944$

返回

差异系数计算实例

例1：某大学一次体检，学生的身高、体重资料如下，问该校学生的身体与体重哪个离散程度大。

	均数	标准差
身高	170 cm	6 cm
体重	60 kg	7 kg

解：$CV_{身高}=\frac{6}{170}\times 100\%=3.5\%$

$CV_{身高}=\frac{7}{60}\times 100\%=11.7\%$

例2：通过同一个测验，一年级学生的平均分数为60分，标准差为4.02分，五年级学生的平均分数为80分，标准差为6.04分，问这两个年级的测验分数中哪一个分数离散程度大。

解：$CV_{一年级}=\frac{4.02}{60}\times 100\%=6.7\%$

$CV_{五年级}=\frac{6.04}{80}\times 100\%=7.55\%$

返回

- 下次上课再见！