集中趋势：数据聚拢位置的衡量

- **均值**
  - **算数平均数**：一组数据中所有数据之和再除以数据的个数

$$A_n = \frac{a_1 + a_2}{n}$$

  - **几何平均数**：n个观察值连乘积的n次方根

$$G_n = \sqrt[n]{a_1 \cdot a_2 \cdot a_3 \cdot \ldots \cdot a_n}$$

  - **调和平均数**：数值倒数的平均数的倒数 `计算结果恒小于等于算术平均数`

$$H_n = \frac{n}{\frac{1}{a_1} + \frac{1}{a_2} + \frac{1}{a_3} + \ldots + \frac{1}{a_n}}$$

    解决在无法掌握总体单位数（频数）的情况下，只有每组的变量值和相应的标志总量，而需要求得平均数的情况下使用的一种数据方法
    用在相同距离但速度不同时，平均速度的计算，相同距离但速度不同时，平均速度的计算，前半段时速60公里，后半段时速30公里〔两段距离相等〕，则其平均速度为两者的调和平均数时速40公里

  - **加权平均数**：把原始数据按照合理的比例

$$\overline{X} = \frac{x_1 f_1 + x_2 f_2 + \ldots + x_k f_k}{n} \text{其中} f_1\text{、}f_2\text{、}\ldots f_k \text{是} x_1\text{、}x_2\text{、}\ldots x_k \text{的权}$$

  - **平方平均数(均方根)**：n个数据的平方的算术平均数的算术平方根

$$M_n = \sqrt{\frac{a_1^2 + a_2^2 + a_3^2 + \ldots + a_n^2}{n}}$$

    分析噪声，也是定义AC波的有效电压或电流的一种最普遍的数学方法

  - 指数平均数[EXPMA]：一种趋向类指标,指数平均数指标是以指数式递减加权的移动平均，对股票收盘价进行算术平均，并根据计算结果来进行分析，用于判断价格未来走势的变动趋势

- **中位数**

  通过把所有观察值高低排序后找出正中间的一个作为中位数

- **众数**

  一组数据中出现次数最多的数值 离散值
  用众数代表一组数据，可靠性较差
  当数值或被观察者没有明显次序（常发生于非数值性资料）时特别有用

- **分位数：**

  将数据按大小排列，最常用到的是四分位数
  Q1 = (N + 1) * 0.25

Q2 = (N + 1) * 0.5

Q3 = (N + 1) * 0.75

平均数、中位数和众数都是来刻画数据平均水平的统计量，中位数刻画了一组数据的中等水平，众数刻画了一组数据中出现次数最多的情况

离中趋势：数据离散程度的衡量

- **标准差(又名 均方差)**

$$\sigma=\sqrt{\frac{1}{N}\sum_{i=1}^{N}(x_i-\mu)^2}$$

值越大 数据越离散 值越小 数据越聚拢

$-1\sigma \sim 1\sigma\ 69\% \quad -1.96\sigma \sim 1.96\sigma\ 95\% \quad -2.58\sigma \sim 2.58\sigma\ 99\%$

数据分布：偏态系数与峰度

- **偏态系数(又名 偏差系数)**
  以平均值与中位数之差对标准差之比率

$$S=\frac{\frac{1}{n}\sum_{i=1}^{n}(x_i-\overline{x})^3}{\left(\frac{1}{n}\sum_{i=1}^{n}(x_i-\overline{x})^2\right)^{\frac{3}{2}}}$$

  - 差异系数(又名 变差系数、离散系数、变异系数)
    数据的标准差与其均值的百分比，是测算数据离散程度的相对指标

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

- **峰态系数**
  数据分布集中强度的衡量

$$K=\frac{\frac{1}{n}\sum_{i=1}^{n}(x_i-\overline{x})^4}{\left(\frac{1}{n}\sum_{i=1}^{n}(x_i-\overline{x})^2\right)^2}$$

- **期望**
  数学期望(mean)（或均值，亦简称期望）是试验中每次可能结果的概率乘以其结果的总和
  随着重复次数接近无穷大，数值的算术平均值几乎肯定地收敛于期望值

$$E(x)=\sum_{k=1}^{\infty}x_k p_k$$

- 某城市有10万个家庭，没有孩子的家庭有1000个，有一个孩子的家庭有9万个，有两个孩子的家庭有6000个，有3个孩子的家庭有3000个。
  则此城市中任一个家庭中孩子的数目是一个随机变量，记为X。它可取值0，1，2，3。
  其中，X取0的概率为0.01，取1的概率为0.9，取2的概率为0.06，取3的概率为0.03。
  则，它的数学期望$E(X) = 0 \times 0.01 + 1 \times 0.9 + 2 \times 0.06 + 3 \times 0.03 = 1.11$，
  即此城市一个家庭平均有小孩1.11个，当然人不可能用1.11个来算，约等于2个。

- **方差**
  概率论中方差用来度量随机变量和其数学期望（即均值）之间的偏离程度
  统计中的方差（样本方差）是每个样本值与全体样本值的平均数之差的平方值的平均数

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

  σ^2为总体方差，X为变量，μ为总体均值，N为总体例数。
  刻画了随机变量的取值对于其数学期望的离散程度

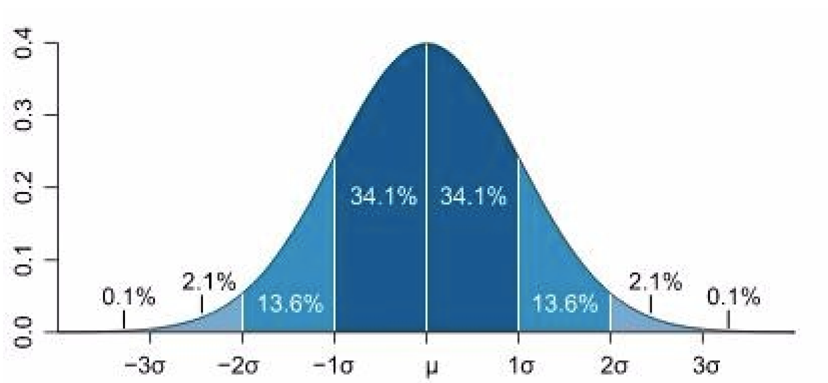
数据分布：分布概率

> 甲乙两个人赌博，他们两人获胜的机率相等，比赛规则是先胜三局者为赢家，赢家可以获得100法郎的奖励。
> 当比赛进行到第四局的时候，甲胜了两局，乙胜了一局，这时由于某些原因中止了比赛，那么如何分配这100法郎才算比较公平?

甲获胜就有两种情况：①甲赢了第四局，比赛结束；②甲输掉了第四局而赢了第五局。于是有，概率P(甲)=1/2+(1/2)*(1/2)=3/4。
而乙获胜的情况就只有一种，同时赢下第四局和第五局，那么，概率P(乙)=(1/2)(1/2)=1/4。
因此，这100法郎就应该分给甲100*3/4=75法郎，分给乙100*1/4=25法郎。
伯努利：**大数定律**

- **正态分布**

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



- **置信区间**
  置信区间是指由样本统计量所构造的总体参数的估计区间，
  置信度也称为可靠度，或置信水平、置信系数，即在抽样对总体参数作出估计时，由于样本的随机性，

其结论总是不确定的。
https://blog.csdn.net/yimingsilence/article/details/78084810

- **三大分布**
  - 卡方分布
    若n个相互独立的随机变量ξ_1，ξ_2，...,ξn，均服从标准正态分布（也称独立同分布于标准正态分布），则这n个服从标准正态分布的随机变量的平方和构成一新的随机变量，其分布规律称为卡方分布
    https://www.cnblogs.com/think-and-do/p/6509239.html
  - T分布
  - F分布