**mean，均值**

**variance，方差**

**standard deviation，标准差**

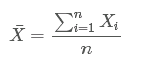
**covariance，协方差**

**covariance matrix，协方差矩阵**

**样本方差:除以（n-1）是关于总体方差的无偏估计量**

**总体方差:**

1. **均值：**



其中为样本均值，X是样本变量，n为样本总数

1. **总体方差：**

方差衡量随机变量和平均值（数学期望）之间的偏离程度。

例子：方差A >方差B， A偏离A的均值比B偏离B的均值要大



σ^2：总体方差

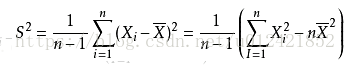
X：变量

N：总数量

μ：均值

1. **样本方差**

**（无偏估计的计算公式）：**



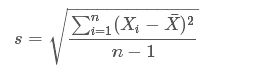
S^2: 为样本方差，X为样本变量，https://img-blog.csdn.net/20180528101456423?watermark/2/text/aHR0cHM6Ly9ibG9nLmNzZG4ubmV0L3UwMTI0MjE4NTI=/font/5a6L5L2T/fontsize/400/fill/I0JBQkFCMA==/dissolve/70为样本均值，n为样本总数

**（样本中心化的公式）**



1. **样本标准差**

**（样本中心化的公式）：**



S：样本标准差，X：样本变量，为样本均值，n为样本总数。除以n-1而不是除以n，为了更好的接近总体的标准差（无偏估计）。方差=（标准差）^2

**例子：（**标准差和方差一般是用来描述一维数据的，在描述两个样本之间的关系，就需要用到协方差了。**）**

**A: [2,4,6]**

**B: [3,4,5]**

A方差：8/3 B 方差：2/3

A均值：4 B 均值：4

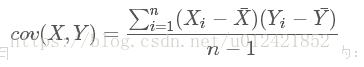
**A的偏离程度大于B的偏离程度！**

1. **协方差**

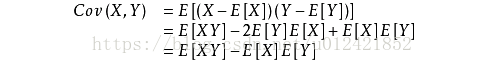
协方差用来衡量两个变量的总体误差。

方差（特殊的协方差）：即当两个变量是相同的情况。

**基于样本集的中心化的协方差公式：**



**E(X)和E(Y)的两个实随机变量X与Y之间的协方差Cov(X,Y)定义为：**



**协方差表示的是两个变量总体误差的期望**

**协方差正值：**两个变量变化趋势相同

**协方差负值：**两个变量变化趋势相反

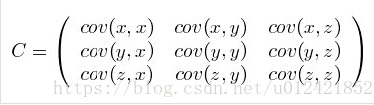
如果X与Y是统计独立的，二者之间的协方差就是0，因为两个独立的随机变量满足E[XY]=E[X]E[Y]。

但是，反过来并不成立。如果X与Y的协方差为0，二者并不一定是统计独立的。

协方差为0的两个随机变量称为是不相关的。

1. **协方差矩阵**

协方差只能处理二维问题，处理多维问题的时候，需要协方差矩阵。



协方差有如下两个公式成立：

Cov (X, X) =Var (X)

Cov (X, Y) =Cov (Y, X)

**特性：**

（1）**协方差矩阵是一个对称阵**

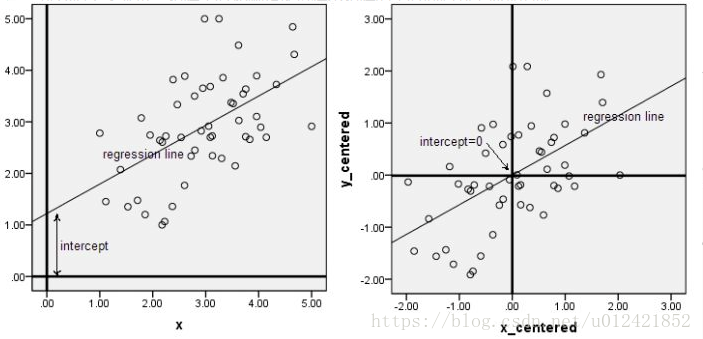
（2）**协方差矩阵的对角线上的元素是每个维度的方差**

（3）**协方差矩阵计算的是一个样本中不同维度之间的协方差，而不是两个或多个样本之间的协方差**

（4）**此处所说的维度就是样例的一个特征，比如一个样本是10\*3大小，则表示有10个样例，每个样例有3个特征**

**7.矩阵中心化**

**样本矩阵中心化：**每一维度减去该维度的均值，使每一维的均值为0；（样本矩阵的中心回到原点）



说明：矩阵中心化的几何意义，就是讲样本集的中心平移到坐标系的原点O上。

（1）数据进行中心化预处理，这样做的目的是要增加基向量的正交性。

（2）数据标准化的目的是消除特征之间的差异性。