## 均值： 均值描述的是样本集合的中间点。

## 标准差：标准差给我们描述的则是样本集合的各个样本点到均值的距离之平均。

## 方差：方差则仅仅是标准差的平方。

## 协方差：如果结果为正值，则说明两者是正相关的(从协方差可以引出“相关系数”的定义)，如果结果为正值，则说明两者是正相关的(从协方差可以引出“相关系数”的定义)，如果为0，也是就是统计上说的“相互独立”。

## 协方差性质：注：

**注：协方差只能处理二维问题。多维问题使用协方差矩阵。**

## 协方差矩阵：

## 三维协方差矩阵：协方差矩阵是一个对称的矩阵，而且对角线是各个维度上的方差。

PCA降维步骤：

**设有m条n维数据**

**1）将原始数据按列组成n行m列矩阵X；**

**2）将X的每一行（代表一个属性字段）进行零均值化，即减去这一行的均值；**

**3）求出协方差矩阵；**

**4）求出协方差矩阵的特征值及对应的特征向量r；**

**5）将特征向量按对应特征值大小从上到下按行排列成矩阵，取前k行组成矩阵P；**

**6）用中心化后的数据矩阵乘以K个特征向量组成的矩阵，得到的矩阵即降到K维的数据。**

## ****维数K的选择：****

**使用一个公式：，表示压缩之后的误差，m所有特征的个数，然后确定一个阈值x，比如0.01，选取一个K,使得error<x,则我们认为这个m可以接受，否则尝试其他。**

## 特征向量和特征值：

**（可以直观的理解：“特征向量是坐标轴，特征值是坐标”）   
向量是具有大小（magnitude）和方向（direction）的几何概念。 特征向量是由满足如下公式的矩阵得到的一个非零向量：**,**其中，****→是特征向量，AA是方阵，λλ是特征值。经过AA变换之后，特征向量的方向保持不变，只是其大小发生了特征值倍数的变化。也就是说，一个特征向量左乘一个矩阵之后等于等比例放缩（scaling）特征向量。**

**矩阵的特征向量是属于并描述数据集结构的向量。   
特征向量和特征值只能由方阵得出，且并非所有方阵都有特征向量和特征值。如果一个矩阵有特征向量和特征值，那么它的每个维度都有一对特征向量和特征值。矩阵的主成分是由其协方差矩阵的特征向量，按照对应的特征值大小排序得到的。最大的特征值就是第一主成分，第二大的特征值就是第二主成分，以此类推。**