https://baike.so.com/doc/5397711-5635055.html

协方差

在概率论和统计学中，协方差用于衡量两个变量的总体误差。而方差是协方差的一种特殊情况，即当两个变量是相同的情况。

期望值分别为E(X) = μ 与 E(Y) = v 的两个实数随机变量X与Y之间的协方差定义为：

COV(X，Y)=E[(X-E(X))(Y-E(Y))]

其中，E是期望值。它也可以表示为：

直观上来看，协方差表示的是两个变量总体误差的方差，这与只表示一个变量误差的方差不同。

如果两个变量的变化趋势一致，也就是说如果其中一个大于自身的期望值，另外一个也大于自身的期望值，那么两个变量之间的协方差就是正值。

如果两个变量的变化趋势相反，即其中一个大于自身的期望值，另外一个却小于自身的期望值，那么两个变量之间的协方差就是负值。

如果X与Y是统计独立的，那么二者之间的协方差就是0。

但是，反过来并不成立。即如果X与Y的协方差为0，二者并不一定是统计独立的。

协方差cov(X,Y)的度量单位是X的协方差乘以Y的协方差。而取决于协方差的相关性，是一个衡量线性独立的无量纲的数。

协方差为0的两个随机变量称为是不相关的。

协方差的意义和计算公式

协方差的意义和计算公式

学过概率统计的孩子都知道，统计里最基本的概念就是样本的均值，方差，或者再加个标准差。首先我们给你一个含有n个样本的集合，依次给出这些概念的公式描述，这些高中学过数学的孩子都应该知道吧，一带而过。

均值： $\bar{X} = \frac{\sum_{i=1}^{n} X_i}{n}$

标准差： $s = \sqrt{\frac{\sum_{i=1}^{n} (X_i - \bar{X})^2}{n-1}}$

方差： $s^2 = \frac{\sum_{i=1}^{n} (X_i - \bar{X})^2}{n-1}$

很显然，均值描述的是样本集合的中间点，它告诉我们的信息是很有限的，而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。以这两个集合为例，[0，8，12，

20]和[8，9，11，12]，两个集合的均值都是10，但显然两个集合差别是很大的，计算两者的标准差，前者是8.3，后者是1.8，显然后者较为集中，故其标准差小一些，标准差描述的就是这种"散布度"。之所以除以n-1而不是除以n，是因为这样能使我们以较小的样本集更好的逼近总体的标准差，即统计上所谓的"无偏估计"。而方差则仅仅是标准差的平方。

为什么需要协方差?

上面几个统计量看似已经描述的差不多了，但我们应该注意到，标准差和方差一般是用来描述一维数据的，但现实生活我们常常遇到含有多维数据的数据集，最简单的大家上学时免不了要统计多个学科的考试成绩。面对这样的数据集，我们当然可以按照每一维独立的计算其方差，但是通常我们还想了解更多，比如，一个男孩子的猥琐程度跟他受女孩子欢迎程度是否存在一些联系啊，嘿嘿~协方差就是这样一种用来度量两个随机变量关系的统计量，我们可以仿照方差的定义：

$$var(X) = \frac{\sum_{i=1}^{n}(X_i - \bar{X})(X_i - \bar{X})}{n-1}$$

来度量各个维度偏离其均值的程度，标准差可以这么来定义：

$$cov(X,Y) = \frac{\sum_{i=1}^{n}(X_i - \bar{X})(Y_i - \bar{Y})}{n-1}$$

协方差的结果有什么意义呢？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)，也就是说一个人越猥琐就越受女孩子欢迎，嘿嘿，那必须的~结果为负值就说明负相关的，越猥琐女孩子越讨厌，可能吗？如果为0，也是就是统计上说的"相互独立"。

从协方差的定义上我们也可以看出一些显而易见的性质，如：

$$1.cov(X,X) = var(X)$$
$$2.cov(X,Y) = cov(Y,X)$$

协方差多了就是协方差矩阵

上一节提到的猥琐和受欢迎的问题是典型二维问题，而协方差也只能处理二维问题，那维数多了自然就需要计算多个协方差，比如n维的数据集就需要计算 n! / ((n-2)!*2) 个协方差，那自然而然的我们会想到使用矩阵来组织这些数据。给出协方差矩阵的定义：

$$C_{n\times n} = (c_{i,j}, c_{i,j} = cov(Dim_i, Dim_j))$$

这个定义还是很容易理解的，我们可以举一个简单的三维的例子，假设数据集有三个维度，则协方差矩阵为

$$C = \begin{pmatrix} cov(x,x) & cov(x,y) & cov(x,z) \\ cov(y,x) & cov(y,y) & cov(y,z) \\ cov(z,x) & cov(z,y) & cov(z,z) \end{pmatrix}$$

可见，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。

Matlab协方差实战

上面涉及的内容都比较容易，协方差矩阵似乎也很简单，但实战起来就很容易让人迷茫了。必须要明确一点，**协方差矩阵计算的是不同维度之间的协方差，而不是不同样本之间的。**这个我将结合下面的例子说明，以下的演示将使用Matlab，为了说明计算原理，不直接调用Matlab的cov函数(蓝色部分为Matlab代码)。

首先，随机产生一个10*3维的整数矩阵作为样本集，10为样本的个数，3为样本的维数。

```
mysample = fix(rand(10,3)*50)
```

```
mysample =

    47    30     2
    11    39    17
    30    46    40
    24    36     0
    44     8     6
    38    20    10
    22    46     9
     0    45    30
    41    20    13
    22    44     9
```

根据公式，计算协方差需要计算均值，那是按行计算均值还是按列呢，我一开始就老是困扰这个问题。前面我们也特别强调了，协方差矩阵是计算不同维度间的协方差，要时刻牢记这一点。样本矩阵的每行是一个样本，每列为一个维度，所以我们要**按列计算均值**。为了描述方便，我们先将三个维度的数据分别赋值：

```
>> dim1 = mysample(:,1);
>> dim2 = mysample(:,2);
>> dim3 = mysample(:,3);
```

计算dim1与dim2，dim1与dim3，dim2与dim3的协方差：

```
>> sum((dim1 - mean(dim1)) .* (dim2 - mean(dim2))) / (size(mysample, 1) - 1)   %
```

得到 -147.0667

```
>> sum((dim1 - mean(dim1)) .* (dim3 - mean(dim3))) / (size(mysample, 1) - 1)   %
```

得到 -82.2667

```
>> sum((dim2 - mean(dim2)) .* (dim3 - mean(dim3))) / (size(mysample, 1) - 1)   %
```

得到 76.5111

搞清楚了这个后面就容易多了，协方差矩阵的对角线就是各个维度上的方差，下面我们依次计算：

```
>> var(dim1)   %得到 227.8778
>> var(dim2)   %得到 179.8222
>> var(dim3)   %得到 156.7111
```

这样，我们就得到了计算协方差矩阵所需要的所有数据，调用Matlab自带的cov函数进行验证：

```
>> cov(mysample)

  227.8778 -147.0667  -82.2667
 -147.0667  179.8222   76.5111
  -82.2667   76.5111  156.7111
```

把我们计算的数据对号入座，是不是一摸一样？

Update

今天突然发现，原来协方差矩阵还可以这样计算，先让样本矩阵中心化，即每一维度减去该维度的均值，使每一维度上的均值为0，然后直接用新的到的样本矩阵乘上它的转置，然后除以(N-1)即可。其实这种方法也是由前面的公式推导而来，只不过理解起来不是很直观，但在抽象的公式推导时还是很常用的！同样给出Matlab代码实现：

```
>> temp = mysample - repmat(mean(mysample), 10, 1);
>> result = temp' * temp ./ (size(mysample, 1) - 1)
result =

  227.8778 -147.0667  -82.2667
 -147.0667  179.8222   76.5111
  -82.2667   76.5111  156.7111
```

总结

理解协方差矩阵的关键就在于牢记它计算的是不同维度之间的协方差，而不是不同样本之间，拿到一个样本矩阵，我们最先要明确的就是一行是一个样本还是一个维度，心中明确这个整个计算过程就会顺流而下，这么一来就不会迷茫了~