**通俗解释协方差与相关系数**

2018-08-22 23:48:00 陈少钦

原文：https://mp.weixin.qq.com/s/46NrpIako2lJ2ZitAQs8Sw

什么是协方差（Covariance）？

协方差表示的是两个变量的总体的误差，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。

以上是某百科的解释。等等！是不是还是觉得比较晦涩难懂呢？对于非理工科的小白来说，如何清晰、形象地理解协方差和相关系数的数学概念呢？没关系，今天红色石头就通过形象生动的例子，通俗易懂地给大家来讲一讲协方差与相关系数。

**1、协方差是怎么来的？**

简单地来说，**协方差**就是反映两个变量 X 和 Y 的相互关系。这种相互关系大致分为三种：**正相关**、**负相关**、**不相关**。

什么是正相关呢？例如房屋面积（X）越大，房屋总价（Y）越高，则房屋面积与房屋总价是正相关的；

什么是负相关呢？例如一个学生打游戏的时间（X）越多，学习成绩（Y）越差，则打游戏时间与学习成绩是负相关的；

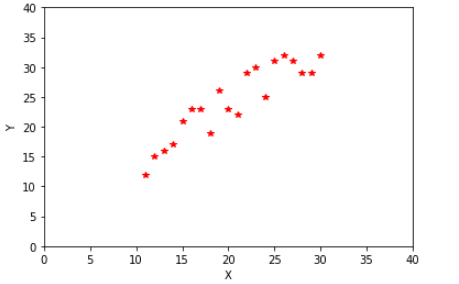
什么是不相关呢？例如一个人皮肤的黑白程度（X）与他的身体健康程度（Y）并无明显关系，所以是不相关的。

我们先来看第一种情况，令变量 X 和变量 Y 分别为：

X = [11 12 13 14 15 16 17 18 19 20 21 22 23 24 25 26 27 28 29 30]

Y = [12 15 17 21 22 21 18 23 26 25 22 28 24 28 30 33 28 34 36 35]

在坐标上描绘出 X 和 Y 的联合分布：



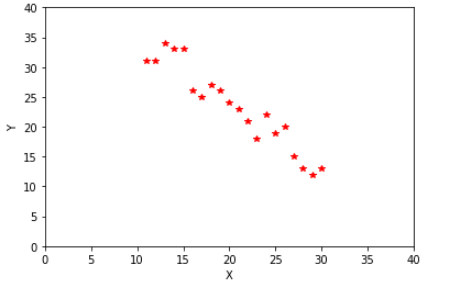
显然，Y 在整体趋势上是随着 X 的增加而增加的，即 Y 与 X 的变化是同向的。这种情况，我们就称 X 与 Y 是**正相关**的。

我们再来看第二种情况，令变量 X 和变量 Y 分别为：

X = [11 12 13 14 15 16 17 18 19 20 21 22 23 24 25 26 27 28 29 30]

Y = [35 35 29 29 28 28 27 26 26 23 21 22 25 19 16 19 20 16 15 16]

在坐标上描绘出 X 和 Y 的联合分布：



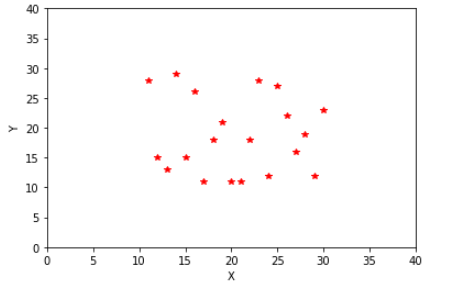
显然，Y 在整体趋势上是随着 X 的增加而减少的，即 Y 与 X 的变化是反向的。这种情况，我们就称 X 与 Y 是**负相关**的。

我们再来看第三种情况，令变量 X 和变量 Y 分别为：

X = [11 12 13 14 15 16 17 18 19 20 21 22 23 24 25 26 27 28 29 30]

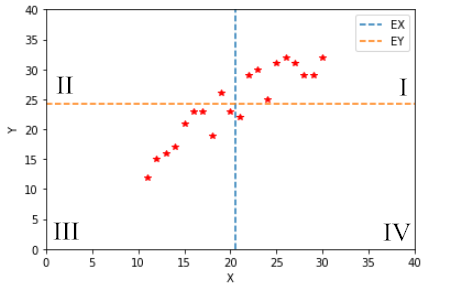
Y = [16 16 28 17 20 26 20 17 21 15 12 29 24 25 16 15 21 13 17 25]

在坐标上描绘出 X 和 Y 的联合分布：



显然，Y 在整体趋势上与 X 的并无正相关或者负相关的关系。这种情况，我们就称 X 与 Y 是**不相关**的。

回过头来，我们来看 X 与 Y 正相关的情况，令 EX、EY 分别是 X 和 Y 的**期望**值。什么是期望呢？在这里我们可以把它看成是**平均值**，即 EX 是变量 X 的平均值，EY 是变量 Y 的平均值。把 EX 和 EY 在图中表示出来得到下面的图形：



上图中，整个区域被 EX 和 EY 分割成 I、II、III、IV 四个区域，且 X 和 Y 大部分分布在 I、III 区域内，只有少部分分布在 II、IV 区域内。

在区域 I 中，满足 X>EX，Y>EY，则有 (X-EX)(Y-EY)>0；

在区域 II 中，满足 X<EX，Y>EY，则有 (X-EX)(Y-EY)<0；

在区域 III 中，满足 X<EX，Y<EY，则有 (X-EX)(Y-EY)>0；

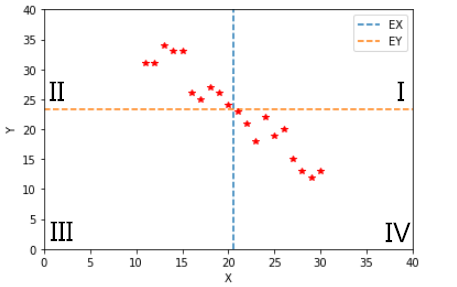
在区域 IV 中，满足 X>EX，Y<EY，则有 (X-EX)(Y-EY)<0。

显然，在区域 I、III 中，(X-EX)(Y-EY)>0；在区域 II、IV 中，(X-EX)(Y-EY)<0。而 X 和 Y 正相关时，数据大部分是分布在 I、III 区域内，只有少部分分布在 II、IV 区域。因此，从平均角度来看，正相关满足：

https://mmbiz.qpic.cn/mmbiz_png/hflWRBRSEZ7V317uUUbK1jYMTtFuPF1nTPiaK8ic5HTOslZuAnOSSQaiaFlJTYACARFQ9ZuiacfefxIM1B4gJce6JQ/640?wxfrom=5&wx_lazy=1

上式表示的是 (X-EX)(Y-EY) 的期望大于零，即 (X-EX)(Y-EY) 的平均值大于零。

然后，再来看 X 和 Y 负相关的情况：



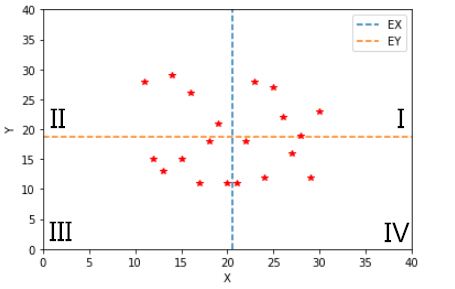
上图中，X 和 Y 大部分分布在 II、IV 区域内，只有少部分分布在 I、III 区域内。

同样，在区域 I、III 中，(X-EX)(Y-EY)>0；在区域 II、IV 中，(X-EX)(Y-EY)<0。而 X 和 Y 负相关时，数据大部分是分布在 II、IV 区域内，只有少部分分布在 I、III 区域。因此，从平均角度来看，负相关满足：

https://mmbiz.qpic.cn/mmbiz_png/hflWRBRSEZ7V317uUUbK1jYMTtFuPF1nt6SRu8KHky8zuibe6t7KtgUGE5rQrHoj2eicK8Gg0HVNXzXibLvvE0FzA/640?wxfrom=5&wx_lazy=1

上式表示的是 (X-EX)(Y-EY) 的期望小于零，即 (X-EX)(Y-EY) 的平均值小于零。

最后，再来看 X 和 Y 不相关的情况：



上图中，X 和 Y 在 I、II、III、IV 区域内近似均匀分布。

同样，在区域 I、III 中，(X-EX)(Y-EY)>0；在区域 II、IV 中，(X-EX)(Y-EY)<0。而 X 和 Y 不相关时，数据在各区域内均匀分布，从平均角度来看，不相关满足：

https://mmbiz.qpic.cn/mmbiz_png/hflWRBRSEZ7V317uUUbK1jYMTtFuPF1nQ6ekqKIHicWDXFHSDzFa9kOaZJcHzcu4L0FLltdbHmzg5Pb14oNaPZQ/640?wxfrom=5&wx_lazy=1

上式表示的是 (X-EX)(Y-EY) 的期望等于零，即 (X-EX)(Y-EY) 的平均值等于零。

综上所述，我们得到以下结论：

* 当 X 和 Y 正相关时：

https://mmbiz.qpic.cn/mmbiz_png/hflWRBRSEZ7V317uUUbK1jYMTtFuPF1nTPiaK8ic5HTOslZuAnOSSQaiaFlJTYACARFQ9ZuiacfefxIM1B4gJce6JQ/640?wxfrom=5&wx_lazy=1

* 当 X 和 Y 负相关时：

https://mmbiz.qpic.cn/mmbiz_png/hflWRBRSEZ7V317uUUbK1jYMTtFuPF1nt6SRu8KHky8zuibe6t7KtgUGE5rQrHoj2eicK8Gg0HVNXzXibLvvE0FzA/640?wxfrom=5&wx_lazy=1

* 当 X 和 Y 不相关时：

https://mmbiz.qpic.cn/mmbiz_png/hflWRBRSEZ7V317uUUbK1jYMTtFuPF1nQ6ekqKIHicWDXFHSDzFa9kOaZJcHzcu4L0FLltdbHmzg5Pb14oNaPZQ/640?wxfrom=5&wx_lazy=1

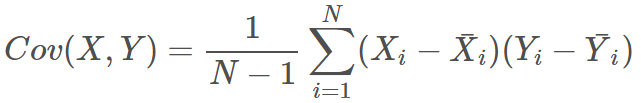
因此，我们就引出了**协方差**的概念，它是表示 X 和 Y 之间相互关系的数字特征。我们定义协方差为：

https://mmbiz.qpic.cn/mmbiz_png/hflWRBRSEZ7V317uUUbK1jYMTtFuPF1nUiaZ2f4BWRASQDUjTcOjcdqyg9uUaYVoZmZAIR07Bo0KWcYSvMOF0Ag/640?wxfrom=5&wx_lazy=1

根据之前讨论的结果，

* **当 Cov(X,Y) > 0 时，X 与 Y 正相关；**
* **当 Cov(X,Y) < 0 时，X 与 Y 负相关；**
* **当 Cov(X,Y) = 0 时，X 与 Y 不相关。**

值得一提的是，**E** 代表求期望值。也可以用平均值来计算协方差：



这里，之所以除以 N-1 而不是 N 的原因是对总体样本期望的无偏估计。顺便提一下，如果令 Y = X，则协方差表示的正是 X 的方差。

下面，我们根据协方差的公式，分别计算上面三种情况下 X 与 Y 的协方差。

X 与 Y 正相关时，Cov(X,Y) = 37.3684；

X 与 Y 负相关时，Cov(X,Y) = -34.0789；

X 与 Y 不相关时，Cov(X,Y) = -1.0263。

**2、相关系数与协方差什么关系？**

我们已经知道了什么是协方差以及协方差公式是怎么来的，如果知道两个变量 X 与 Y 的协方差与零的关系，我们就能推断出 X 与 Y 是正相关、负相关还是不相关。那么有一个问题：协方差数值大小是否代表了相关程度呢？也就是说如果协方差为 100 是否一定比协方差为 10 的正相关性强呢？

请看下面这个例子！

变量 X1 与 Y1 分别为：

X1 = [11 12 13 14 15 16 17 18 19 20 21 22 23 24 25 26 27 28 29 30]

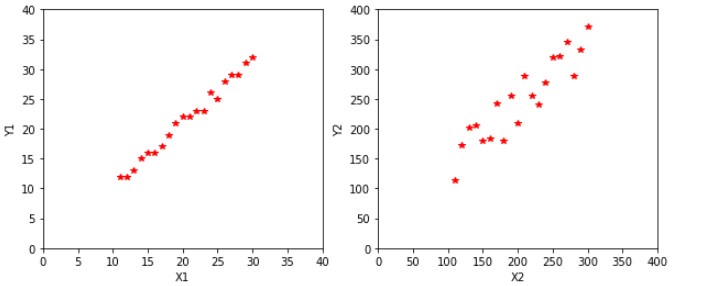
Y1 = [12 12 13 15 16 16 17 19 21 22 22 23 23 26 25 28 29 29 31 32]

变量 X2 和 Y2 分别为：

X2 = [110 120 130 140 150 160 170 180 190 200 210 220 230 240 250 260 270 280 290 300]

Y2 = [113 172 202 206 180 184 242 180 256 209 288 255 240 278 319 322 345 289 333 372]

X1、Y1 和 X2、Y2 分别联合分布图，如下所示：



显然，从图中可以看出，X1、Y1 和 X2、Y2 都呈正相关，而且 X1 与 Y1 正相关的程度明显比 X2 与 Y2 更大一些。接下来，我们计算两幅图的协方差看看是不是这样。

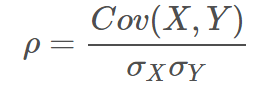
Cov(X1,Y1) = 37.5526

Cov(X2,Y2) = 3730.26

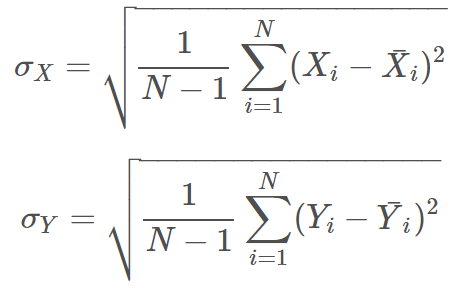
意外！X2 与 Y2 的协方差竟然比 X1 与 Y1 的协方差还大 100 倍。看来并不是协方差越大，正相关程度越高。这到底是为什么呢？

其实，出现这种情况的原因是两种情况数值变化的幅值不同（或者量纲不同）。计算协方差的时候我们并没有把不同变量幅值差异性考虑进来，在比较协方差的时候也就没有一个统一的量纲标准。

所以，为了消除这一影响，为了准确得到变量之间的相似程度，我们需要把协方差除以各自变量的标准差。这样就得到了**相关系数**的表达式：



可见，相关系数就是在协方差的基础上除以变量 X 和 Y 的标准差。其中标准差的计算公式为：



为什么除以各自变量的标准差就能消除幅值影响呢？这是因为标准差本身反映了变量的幅值变化程度，除以标准差正好能起到抵消的作用，让协方差标准化。这样，相关系数的范围就被归一化到 [-1,1] 之间了。

下面，我们就来分别计算上面这个例子中 X1、Y1 和 X2、Y2 的相关系数。

ρ(X1,Y1) = 0.9939

ρ(X2,Y2) = 0.9180

好了，我们得到 X1 与 Y1 的相关系数大于 X2 与 Y2 的相关系数。这符合实际情况。也就是说，根据相关系数，我们就能判定两个变量的相关程度，得到以下结论：

* **相关系数大于零，则表示两个变量正相关，且相关系数越大，正相关性越高；**
* **相关系数小于零，则表示两个变量负相关，且相关系数越小，负相关性越高；**
* **相关系数等于零，则表示两个变量不相关。**

回过头来看一下协方差与相关系数的关系，其实，相关系数是协方差的标准化、归一化形式，消除了量纲、幅值变化不一的影响。实际应用中，在比较不同变量之间相关性时，使用相关系数更为科学和准确。但是协方差在机器学习的很多领域都有应用，而且非常重要！更多协方差的应用红色石头以后会给大家慢慢讲解哦！

**参考文献**：

https://www.cnblogs.com/tsingke/p/6273970.html

https://www.zhihu.com/question/20852004