# 皮尔逊积矩相关系数（Pearson product-moment correlation coefficient）

## 1 定义

**在统计学中，皮尔逊积矩相关系数（Pearson product-moment correlation coefficient），有时也简称为 PMCC，通常用 r 或是 ρ 表示，是用来度量两个变量 X 和 Y 之间的相互关系（线性相关）的，取值范围在[-1,+1]之间。**皮尔逊积矩相关系数在学术研究中被广泛应用来度量两个变量线性相关性的强弱，它是由 Karl Pearson 在 19 世纪 80 年代从 Francis Galton 介绍的想法基础发展起来的，但是发展后原想法相似但略有不同的，这种相关系数常被称为“Pearson 的 r”。

两个变量之间的皮尔逊积矩相关系数定义为这两个变量的协方差与二者标准差积的商，即

$$\rho_{XY} = \frac{\operatorname{cov}(X,Y)}{\sigma_X \sigma_Y} = \frac{E(X-\mu_X)(Y-\mu_Y)}{\sigma_X \sigma_Y}$$

上式定义了总体相关系数，一般用希腊字母 ρ（rho）表示。若用样本计算的协方差和标准差代替总体的协方差和标准差，则为样本相关系数，一般用 r 表示：

$$r = \frac{\sum_{i=1}^{n}(X_i - \bar{X})(Y_i - \bar{Y})}{\sqrt{\sum_{i=1}^{n}(X_i - \bar{X})^2}\sqrt{\sum_{i=1}^{n}(Y_i - \bar{Y})^2}}$$

另外一个与上式等效的定义相关系数的公式是通过标准化以后变量均值的积定义的。假设样本可以记为 $(X_i, Y_i)$，则样本 Pearson 相关系数为

$$r = \frac{1}{n-1}\sum_{i=1}^{n}\left(\frac{X_i - \bar{X}}{s_X}\right)\left(\frac{Y_i - \bar{Y}}{s_Y}\right)$$

其中 $\frac{X_i - \bar{X}}{s_X}$，$\bar{X}$ 和 $s_X$ 分别为标准化变量，样本均值和样本标准差。

## 2 皮尔逊积矩相关系数的数学特性

不论是样本的还是总体的 Pearson 相关系数绝对值均小于等于 1，相关系数等于 1 或-1 时，所有数据的点都精确地落在一条直线上（为样本相关系数的情况），或是两变量的分布完全由一条直线支撑（为总体相关系数的情况）。Pearson 相关系数具有对称性，即：$\operatorname{corr}\operatorname{corr}(X,Y) = \operatorname{corr}(Y,X)$。

Pearson 相关系数的一个关键的特性就是它并不随着变量的位置或是大小的变化而

变化。也就是说，我们可以把 X 变为 a+bX，把 Y 变为 c+dY，其中 a，b，c 和 d 都是常数，而并不会改变相互之间的相关系数（这点对总体和样本 Pearson 相关系数都成立）。

Pearson 相关系数可以用原点矩的形式表示。因为

$$\mu_X = E(X)，\quad \sigma_X^2 = [E(X) - X]^2 = E(X^2) - E^2(X)，$$

对于 Y 也有相似的表达式。又

$$E[(X - E(X))E(Y - E(Y))] = E(XY) - E(X)E(Y)$$

于是式(1)可写为

$$\rho_{XY} = \frac{E(XY) - E(X)E(Y)}{\sqrt{E(X^2) - E^2(X)}\sqrt{E(Y^2) - E^2(Y)}}$$

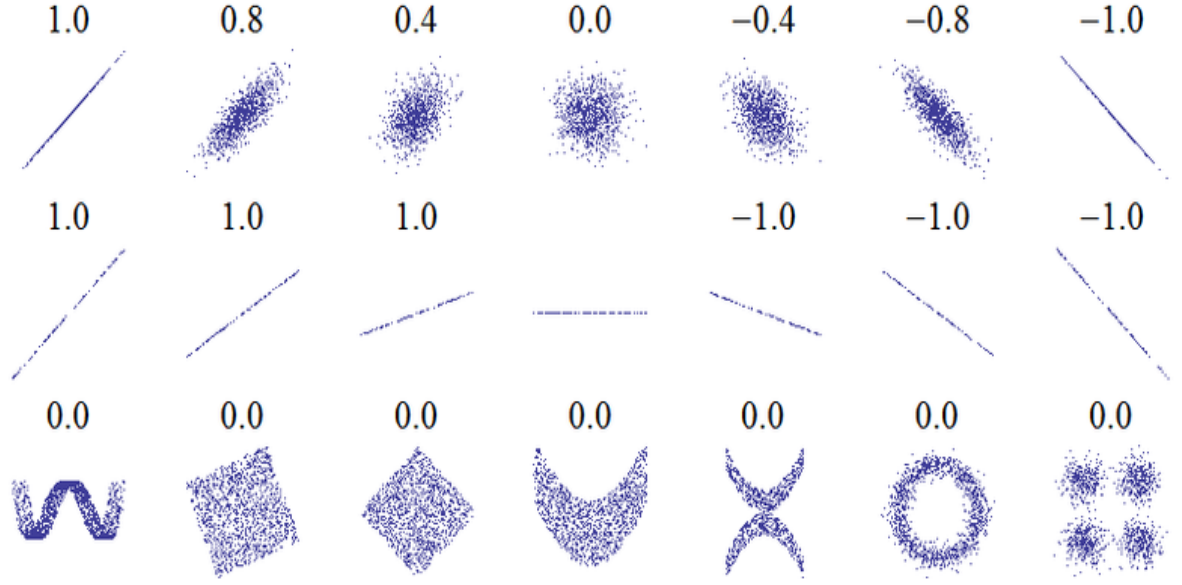
上述形式对于样本的 Pearson 相关系数同样是可用的，有

$$r_{xy} = \frac{\sum x_i y_i - n\bar{x}\bar{y}}{(n-1)s_x s_y} = \frac{n\sum x_i y_i - \sum x_i \sum y_i}{\sqrt{n\sum x_i^2 - (\sum x_i)^2}\sqrt{n\sum y_i^2 - (\sum y_i)^2}}$$

上式提供了一个非常简单的计算样本相关系数的算法，但是有时受数据的影响，可上式可能存在数值上的不稳定性。

相关系数取值范围为[-1,1]。取 1 时表示变量 X 和 Y 之间具有线性变化的关系，即 Y 随着 X 的增加而增加，而且所有的点都落在一条直线上。取-1 时则是所有点落在一条直线上，但是变量 Y 随着 X 的增加而减小。相关系数值为 0 是表示变量之间没有线性相关关系。

更一般地，应该注意到，只要 $X_i$ 和 $Y_i$ 落在各自均值的同一侧，那么 $(X_i - \bar{X})(Y_i - \bar{Y})$ 就是大于 0 的。也就是说，只要 $X_i$ 和 $Y_i$ 同时趋近于大于或是同时趋近于小于他们各自的均值，那么它们的相关系数为正。反之，当二者区于在相反的一边时，二者相关系数为负。




几种的（x，y）点即相应的 x、y 的相关系数。可以看出，相关反映线性关系分散程度和方向（第一行），但是不能反映线性关系时的斜率（第二行），也不能反映出非线性关系的许多方面（最底下一行）。注：图中第二行第四个小图的直线斜率是 0，在这种情况下，相关系数是没有意义的，因为 Y 的方差是零。

# 3 几何解释

对于相对中心性的数据（例如，一组已经通过样本均值转换为均值为 0 的数据），相关系数可以看做是由两随机变量样本绘出的两个向量之间夹角的余弦值。

有些学者则比较倾向于非中心性（费皮尔逊兼容）的相关系数。以下通过一个例子比较二者之间的差异。

假设有 5 个国家，国民生产总值分别为 10 亿美元、20 亿美元、30 亿美元、50 亿美元和 80 亿美元，而贫困人数占总人口的比例分别为 11%、12%、13%、15%和 18%。则可令 X = （10，20，30，50，80），Y = （0.11，0.12，0.13，0.15，0.18）。

有一般的计算两个向量之间的角度的过程（点乘）可得非中心性相关系数为：

$$\cos\theta = \frac{x \cdot y}{\|x\|\|y\|} = \frac{2.93}{\sqrt{103}\sqrt{0.0983}} = 0.920814711$$

应该注意到，上述数据是特意从完全线性相关的线性函数 Y=0.10+0.001X 中挑选出来的，所以 Pearson 相关系数应该精确地为 1。将数据中心化（将 X 减去 E(X)=38，Y 减去 E(Y)=0.138），可得 X’=（-28，-18，-0.8，12，42），Y’=（-0.028，-0.018，-0.08，0.012，0.042），并有

$$\cos\theta' = \frac{x' \cdot y'}{\|x'\|\|y'\|} = \frac{3.08}{\sqrt{3080}\sqrt{0.00308}} = 1 = \rho_{xy}$$

跟期望的一样。

相关系数大小与相关性大小的关系

许多学者都提出了通过相关系数大小判断变量相关性的标准。但是正如 Cohen（1988）所指出的一样，这些标准或多或少的有些武断，不应该过于严格地遵守。相同相关系数对相关性大小的判断取决于不同的背景和目的。同样是 0.9 的相关系数，在使用很精确的仪器验证物理定律的时候可能被认为是很低的，但是社会科学中，在评定许多复杂因素的贡献时，却可能被认为是很高的相关性。

相关系数与相关性的关系

| 相关性 | 负值 | 正值 |
|---|---|---|
| 不相关 | −0.09～0.0 | 0.0～0.09 |
| 低相关 | −0.3～−0.1 | 0.1～0.3 |
| 中等相关 | −0.5～−0.3 | 0.3～0.5 |
| 显著相关 | −1.0～−0.5 | 0.5～1.0 |

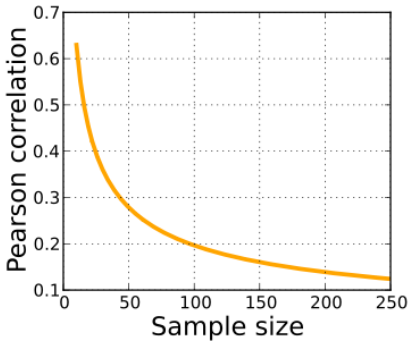
# 4 对数据分布的敏感性

## 4.1 存在性

总体的 Pearson 相关系数是通过原点矩来定义的，所以二元概率分布的总体协方差以及变量边缘总体反差必须是有意义且是非零的。一些概率分布例如柯西（Cauchy）分布的反差就是无意义的，因此在 X 或 Y 服从这种分布时，ρ也是没有意义的。在一些实际应用中，例如那些涉及数据在尾部比较集中的情况，考虑这点就是很重要的。但是，相关系数的存在性通常不是我们关注的焦点，因为一般只要分布是有界的，那么ρ就可以被定义。

## 4.2 大样本性

在二元正态分布中，若已知变量的边缘分布的均值和标准差，那么由 Pearson 相关系数就可以完全确定该分布的特性。但是对于其它的二元分布，情况就有所不同。然而，不论变量之间的联合概率密度函数是不是正态的，Pearson 相关系数都是用来衡量两个随机变量之间的线性相关程度的。对于二元正态数据，样本的相关系数是总体相关系数的极大似然估计，并且具有渐进无偏性和有效性，也即是说在数据来自正态分布，且样本大小适中或是足够大的时候，不可能构造一个比样本相关相关系数更加精确的量来估计变量之间的相关性。对于非正态总体，样本相关系数依然是渐进无偏的，但是可能不是有效的估计。只要样本均值、方差、协方差是一致的（可以通过应用大数定律来保证），样本相关系数是总体相关系数的一个一致估计量。




图中显示了在给定的样本大小时，在置信水平为 0.05 时，具有显著非零 Pearson 相关系数的的最小值。A graph showing the minimum value of Pearson's correlation coefficient that is significantly different from zero at the 0.05 level, for a given sample size.

# 5 鲁棒性（Robustness）

与其他一些广泛应用的统计量相同，样本统计量 r 是不可靠的，在存在异常值的时候，r 的值可能会误导我们。也就是说，PMCC 不仅受变量分布的影响，还随异常值非常敏感。观察 X、Y 之间的散点图，就可以看出，缺少鲁棒性确实是一个很大的问题，在这种情况下，就需要采用更加稳健的参量来度量变量的相关性。但是值得一提的是，无论采用多么稳健的参量来度量变量之间的相关性，都与 Pearson 相关系数在数值大小保持很好的一致性。

基于 Pearson 相关系数的统计推断对数据的分布类型是很敏感的。**所以只有在数据是近似正态分布的时候，基于 Fisher 变换的精确检验和近似检验才能被采用，否则就可能导致错误的结论。**在某些情况下，引导可用于构造置信区间，并置换测试可用于进行假设检验。在二元正态不成立时，非参数的方法在某些情况下可能会得到更有意义的结果。但这些方法的标准版本依赖于数据的互换性，也就是说，在没有特定的顺序或是数据可供分析时，可能影响相关估计的行为。

# Spearman 秩相关系数（Spearman's rank correlation coefficient）

**Pearson 线性相关系数只是许多可能中的一种情况，为了使用 Pearson 线性相关系数必须假设数据是成对地从正态分布中取得的，并且数据至少在逻辑范畴内必须是等间距的数据。如果这两条件不符合，一种可能就是采用 Spearman 秩相关系数来代替 Pearson 线性相关系数。**Spearman 秩相关系数是一个非参数性质（与分布无关）的秩统计参数，由 Spearman 在 1904 年提出，用来度量两个变量之间联系的强弱(Lehmann and D'Abrera 1998)。Spearman 秩相关系数可以用于 R 检验，同样可以在数据的分布使得 Pearson 线性相关系数不能用来描述或是用来描述或导致错误的结论时，作为变量之间单调联系强弱的度量。

在统计学中，Spearman 秩相关系数或称为 Spearman 的 ρ，是由 Charles Spearman 命名的，一般用希腊字母 $\rho_s$（rho）或是 $r_s$ 表示。Spearman 秩相关系数是一个非参数的度量两个变量之间的统计相关性的指标，用来评估当用单调函数来描述是两个变量之间的关系有多好。在没有重复的数据的情况下，如果一个变量是两外一个变量的严格单调的函数，则二者之间的 Spearman 秩相关系数就是+1 或-1，称变量完全 Spearman 相关。

Spearman 秩相关系数通常被认为是排列后的变量之间的 Pearson 线性相关系数，在实际计算中，有更简单的计算 $\rho_s$ 的方法。假设原始的数据 xi，yi 已经按从大到小的顺序排列，记 x’i，y’i 为原 xi，yi 在排列后数据所在的位置，则 x’i，y’i 称为变量 x’i，y’i 的秩次，则 di=x’i-y’i 为 xi，yi 的秩次之差。

如果没有相同的秩次，则 $\rho_s$ 可由下式计算

$$\rho_s = 1 - \frac{6\sum d_i^2}{n(n^2-1)}$$

如果有相同的秩次存在，那么就需要计算秩次之间的 Pearson 的线性相关系数

$$\rho_s = \frac{\sum_i (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_i (x_i - \bar{x})^2 \sum_i (y_i - \bar{y})^2}}$$

一个相同的值在一列数据中必须有相同的秩次，那么在计算中采用的秩次就是数值在按从大到小排列时所在位置的平均值。表 1 为一个球平均秩次的例子。注意在秩次相同时，用他们在排列后的数据中所在的位置的平均值作为秩次。

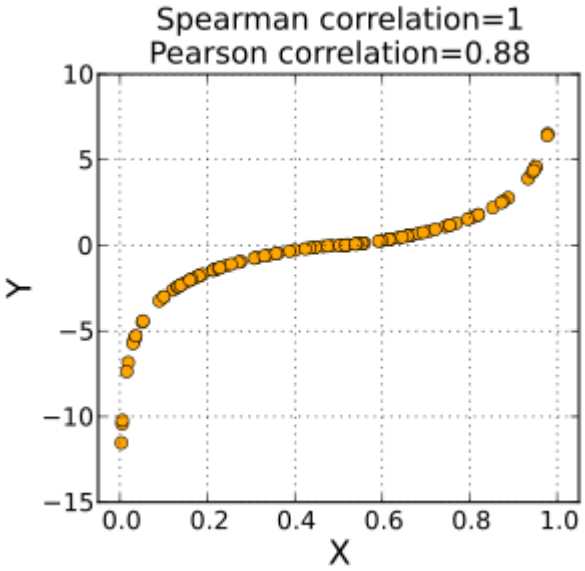
表 1　有相同数值时秩次的计算

| 变量 $x_i$ | 从大到小排列时的位置 | 秩次 $x'_i$ |
|---|---|---|
| 0.8 | 5 | 5 |
| 1.2 | 4 | (4+3)/2=3.5 |
| 1.2 | 3 | (4+3)/2=3.5 |
| 2.3 | 2 | 2 |
| 18 | 1 | 1 |

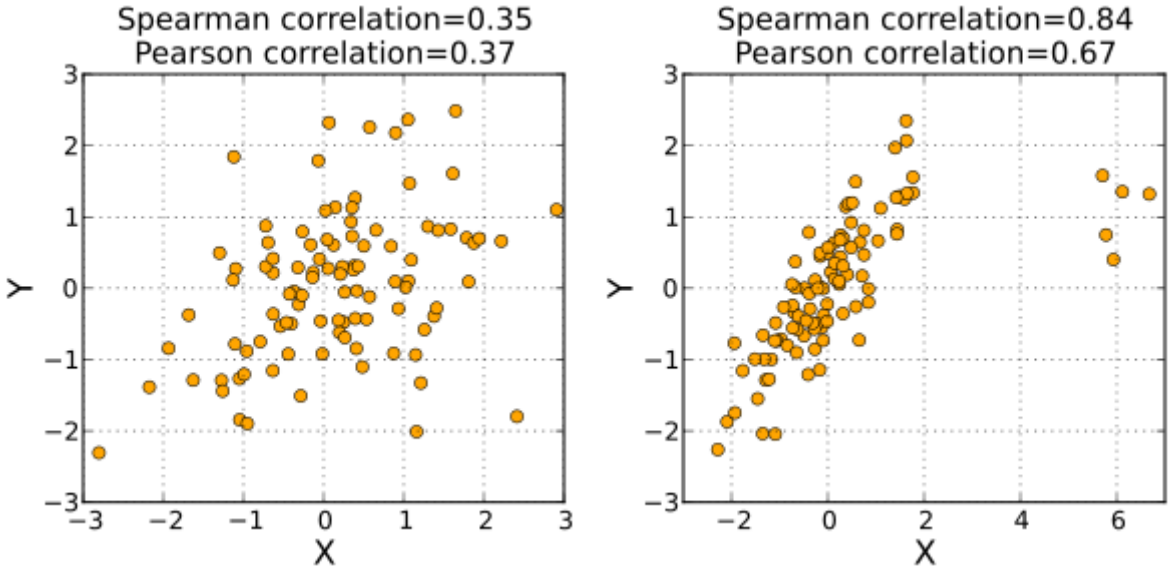
Spearman 秩相关系数的符号表示 X 和 Y 之间联系的方向。如果 Y 随着 X 的增加而增加，那么 Spearman 秩相关系数是正的，反之，若果 Y 随着 X 的增加而减小，Spearman 秩相关系数就是负的。Spearman 秩相关系数为 0 表示随着 X 的增加，Y 没有增大或减小的趋势。随着 X 和 Y 越来越接近严格单调的函数关系，Spearman 秩相关系数在数值上越来越大。当 X、Y 有严格单增的关系是，它们之间的 Spearman 秩相关系数为 1，

反之，在 X、Y 有严格单减的关系时，Spearman 秩相关系数为-1。严格单增的关系为对于任意的两对数据值 Xi，Yi 和 Xj，Yj，Xi-Yi 和 Xj-Yj 都具有相同的符号。严格单减则上述差值在任何时候都具有相反的符号。

**Spearman 秩相关系数经常被称为非参数相关系数**，这具有两层含义：第一，只要在 X 和 Y 具有单调的函数关系的关系，那么 X 和 Y 就是完全 Spearman 相关的，这与 Pearson 相关性不同，后者只有在变量之间具有线性关系时才是完全相关的。另外一个关于 Spearman 秩相关系数的非参数性的理解就是样本之间精确的分布可以在不知道 X 和 Y 的联合概率密度函数时获得。

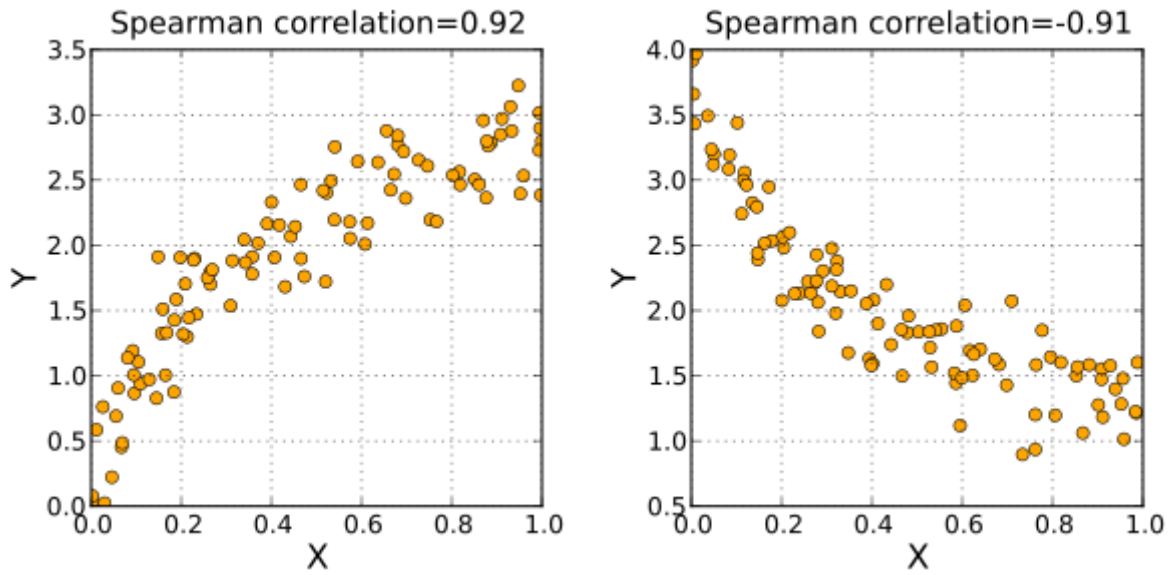



不管变量之间的关系是不是线性的，只要变量之间具有严格的单调增加的函数关系，变量之间的 Spearman 秩相关系数就是 1，相同情况下，Pearson 相关性在变量不是线性函数关系时，并不是完全相关的。




在数据大略地呈椭圆形分布，而且没有明显的外形轮廓的时候，Spearman 秩相关系数和 Pearson 线性相关系数大小比较接近。

Spearman 秩相关系数对样本的尾部与具有明显的外形轮廓样本偏离比较大的情况没有 Pearson 线性相关系数敏感。




正的 Spearman 秩相关系数对应于 X、Y 之间单调增加的变化趋势，负的 Spearman 秩相关系数对应于 X、Y 之间单调减小的变化趋势。