

皮尔逊积矩相关系数

维基百科，自由的百科全书

在统计学中，皮尔逊积矩相关系数（英语：Pearson product-moment correlation coefficient，又称作 **PPMCC** 或 **PCCs**^[1], 文章中常用r或Pearson's r表示）用于度量两个变量X和Y之间的相关程度（线性相关），其值介于-1与1之间。在自然科学领域中，该系数广泛用于度量两个变量之间的线性相关程度。它是由卡尔·皮尔逊从弗朗西斯·高尔顿在19世纪80年代提出的一个相似却又稍有不同的想法演变而来。^{[2][3]}这个相关系数也称作“皮尔森相关系数r”。

目录

定义

数学特性

解释

- 几何学角度的解释
- 对相关系数大小的解释
- 皮尔逊距离

统计推断：显著性检验与置信区间

- 随机采样方法
- 自助抽样法
- 基于数学近似的方法
- 准确服从高斯分布的数据
- 费舍尔变换
- 置信区间

皮尔逊相关系数和最小方差回归分析

数据分布的敏感度

- 存在性
- 大样本的特性
- 稳健性

计算加权相关系数

去除相关性

反射相关性

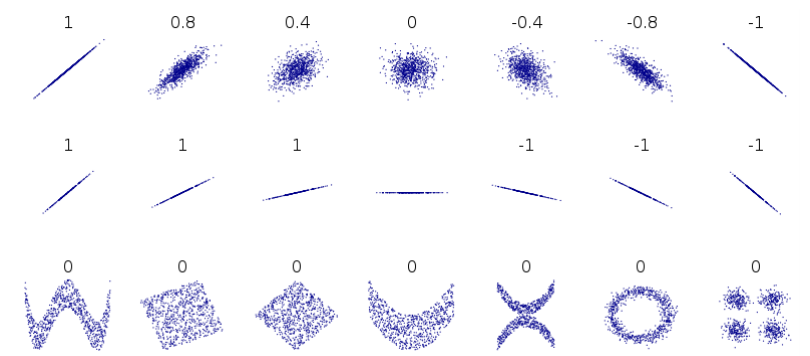
比例关系

强噪声条件下

维基相关条目

外部链接

注脚



几组(x, y)的点集，以及各个点集中x和y之间的相关系数。我们可以发现相关系数反映的是变量之间的线性关系和相关性的方向（第一排），而不是相关性的斜率（中间），也不是各种非线性关系（第三排）。请注意：中间的图中斜率为0，但相关系数是没有意义的，因为此时变量Y是0

定义

两个变量之间的皮尔逊相关系数定义为两个变量之间的协方差和标准差的商：

$$\rho_{X,Y} = \frac{\text{cov}(X,Y)}{\sigma_X \sigma_Y} = \frac{E[(X - \mu_X)(Y - \mu_Y)]}{\sigma_X \sigma_Y}$$

上式定义了总体相关系数，常用希腊小写字母 ρ (rho) 作为代表符号。估算样本的协方差和标准差，可得到样本相关系数(样本皮尔逊系数)，常用英文小写字母 r 代表：

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

r 亦可由 (X_i, Y_i) 样本点的标准分数均值估算，得到与上式等价的表达式：

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

其中 $\frac{X_i - \bar{X}}{\sigma_X}$ 、 \bar{X} 及 σ_X 分别是 X_i 样本的标准分数、样本平均值和样本标准差。

数学特性

总体和样本皮尔逊系数的绝对值小于或等于1。如果样本数据点精确的落在直线上（计算样本皮尔逊系数的情况），或者双变量分布完全在直线上（计算总体皮尔逊系数的情况），则相关系数等于1或-1。皮尔逊系数是对称的： $\text{corr}(X,Y) = \text{corr}(Y,X)$ 。

皮尔逊相关系数有一个重要的数学特性是，因两个变量的位置和尺度的变化并不会引起该系数的改变，即它该变化的不变量(由符号确定)。也就是说，我们如果把X移动到 $a + bX$ 和把Y移动到 $c + dY$ ，其中 a 、 b 、 c 和 d 是常数，并不会改变两个变量的相关系数（该结论在总体和样本皮尔逊相关系数中都成立）。我们发现更一般的线性变换则会改变相关系数：参见之后章节对该特性应用的介绍。

由于 $\mu_X = E(X)$, $\sigma_X^2 = E[(X - E(X))^2] = E(X^2) - E^2(X)$ ，Y也类似, 并且

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

故相关系数也可以表示成

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

对于样本皮尔逊相关系数：

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

以上方程给出了计算样本皮尔逊相关系数简单的单流程算法，但是其依赖于涉及到的数据，有时它可能是数值不稳定的。

解释

皮尔逊相关系数的变化范围为-1到1。系数的值为1意味着X和Y可以很好的由直线方程来描述，所有的数据点都很好的落在一条直线上，且Y随着X的增加而增加。系数的值为-1意味着所有的数据点都落在直线上，且Y随着X的增加而减少。系数的值为0意味着两个变量之间没有线性关系。

更一般的,我们发现，当且仅当 X_i and Y_i 均落在他们各自的均值的同一侧，则 $(X_i - \bar{X})(Y_i - \bar{Y})$ 的值为正。也就是说，如果 X_i 和 Y_i 同时趋向于大于,或同时趋向于小于他们各自的均值，则相关系数为正。如果 X_i 和 Y_i 趋向于落在他们均值的相反一侧，则相关系数为负。

几何学角度的解释

对于没有进行中心化的数据,相关系数与两条可能的回归线 $y=g_x(x)$ 和 $x=g_y(y)$ 夹角的余弦值一致。

对于中心化过的数据 (也就是说,数据移动一个样本平均值以使其均值为0),相关系数也可以被视作由两个随机变量向量夹角 θ 的余弦值 (见下方)。

一些人 倾向于使用非中心化的相关系数 (non-Pearson-compliant)。比较如下。

例如，有5个国家的国民生产总值分别为 10, 20, 30, 50 和 80 亿美元。假设这5个国家 (顺序相同) 的贫困百分比分别为 11%, 12%, 13%, 15%, 和 18%。令 \mathbf{x} 和 \mathbf{y} 分别等于包含上述5个数据的向量: $\mathbf{x} = (1, 2, 3, 5, 8)$ 和 $\mathbf{y} = (0.11, 0.12, 0.13, 0.15, 0.18)$ 。

利用通常的方法计算两个向量之间的夹角 θ (参见 数量积), 未中心化的相关系数是:

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

我们发现以上的数据特意选定为完全相关: $y = 0.10 + 0.01x$ 。于是，皮尔逊相关系数应该等于1。将数据中心化 (通过 $E(\mathbf{x}) = 3.8$ 移动 \mathbf{x} 和通过 $E(\mathbf{y}) = 0.138$ 移动 \mathbf{y}) 得到 $\mathbf{x} = (-2.8, -1.8, -0.8, 1.2, 4.2)$ 和 $\mathbf{y} = (-0.028, -0.018, -0.008, 0.012, 0.042)$, 从中,

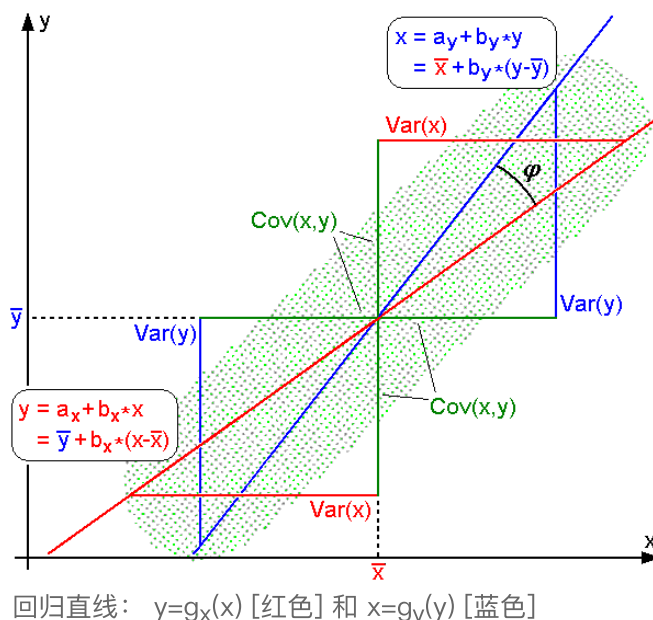
$$\cos \theta = \frac{\mathbf{x} \cdot \mathbf{y}}{\|\mathbf{x}\| \|\mathbf{y}\|} = \frac{0.308}{\sqrt{30.8}\sqrt{0.00308}} = 1 = \rho_{xy},$$

对相关系数大小的解释

一些著作的作者^{[4][5]}给出了某些解释相关系数的指南。然而,所有这些标准从某种意义上说是武断的和不严格的。^[5]对相关系数的解释是依赖于具体的应用背景和目的的。例如，若是在运用高性能的仪器来验证一个物理定律实验这样的应用背景下，0.9的相关系数可能是很低的。但如果是应用在社会科学中，由于社会科学受到各种复杂多变因素影响，0.9的相关系数是相当高的。

皮尔逊距离

皮尔逊距离度量的是两个变量X和Y，它可以根据皮尔逊系数定义成^[6]



相关性	负	正
无	-0.09 to 0.0	0.0 to 0.09
弱	-0.3 to -0.1	0.1 to 0.3
中	-0.5 to -0.3	0.3 to 0.5
强	-1.0 to -0.5	0.5 to 1.0

$$d_{X,Y} = 1 - \rho_{X,Y}.$$

我们可以发现，皮尔逊系数落在 $[-1, 1]$ ，而皮尔逊距离落在 $[0, 2]$ 。

统计推断：显著性检验与置信区间

基于皮尔逊相关系数的统计推断通常关注以下两个目标。

1. 验证零假设是否为真，即相关系数 ρ 是否等于 0，该相关系数使用的是样本相关系数 r 。
2. 在给定的置信水平 α 之下，构建一个围绕 r 的置信区间。

随机采样方法

显著性检验提供了一种假设检验和构造置信区间的直接方法。

对皮尔逊相关系数的显著性检验包括以下两个步骤：

1. 随机地将原始的数据对 (x_i, y_i) 重新定义成数据集 $(x_{i'}, y_{i'})$ ，其中 i' 表示数列 $\{1, \dots, n\}$ 。数列 i' 的选取是随机的，以相同的概率落在 $n!$ 种可能的数列中。这等价于随机地“不可重复地”从数列 $\{1, \dots, n\}$ 中选取 i' 。一种相近的且合乎情理的方法（自助抽样法）是“可重复地”从数列 $\{1, \dots, n\}$ 中选取 i 和 i'
2. 由随机数据构造相关系数 r 。

为了完成显著性检验，需要多次重复步骤(i)和(ii)。显著性检验的P值是由测试数据除以步骤（ii）得到的 r ，其中 r 大于由原始数据计算出的皮尔逊相关系数。在这里“大”可能是绝对值比较大或者是数值比较大，这取决于测试使用的是双边检验或者是单边检验。

自助抽样法

自助抽样法可以被用来构造皮尔逊系数的置信区间。在“非参数”的自助抽样法中，“可重复”地从观测数据集 n 中重新采样 n 对的 (x_i, y_i) 数据，用来计算相关系数 r 。这个过程重复了大量次数，重新采样后数据的 r 值的分布被用来估计统计学上的样本分布。 ρ 的95%的置信区间可以被定义成重新采样样本 r 值的%2.5到%97.5之间。

基于数学近似的方法

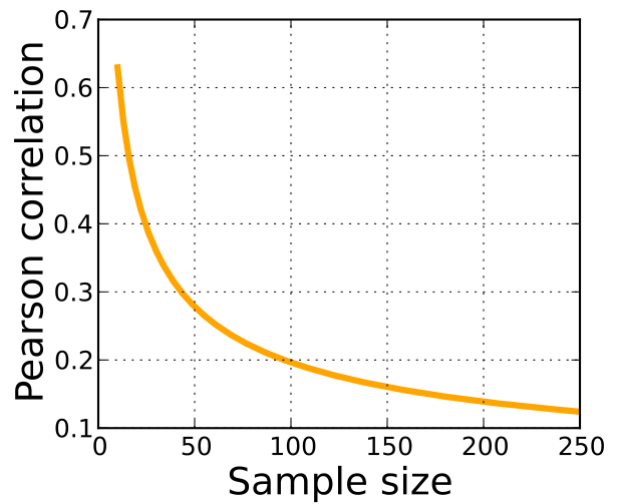
对于近似高斯分布的数据，皮尔逊相关系数的样本分布近似于自由度为 $N - 2$ 的 t 分布。特别地，如果两个变量服从双变量正态分布，变量

$$t = r \sqrt{\frac{n-2}{1-r^2}}$$

也会服从不相关的 t 分布。^[7] 如果样本容量不是特别小，这个结论也大致成立，即便观测数据不是正态分布的。^[8] 如果需要构建置信区间和进行有力的分析，还需要采用如下的可逆变换

$$r = \frac{t}{\sqrt{n-2+t^2}}.$$

或者，也可以采用大量采样数据的方法。



图表显示对于给定的样本大小，在0.05的置信度上，皮尔逊相关系数显著不为零。

早期对样本相关系数的研究得益于R. A. Fisher^{[9][10]}和A. K. Gayen^[11]的工作。另一篇早期的论文^[12]给出了在小样本的情况下总体相关系数 ρ 的图表, 并讨论了相关的计算方法。

准确服从高斯分布的数据

准确的双变量样本相关系数的分布是^{[13][14]}

$$f(r) = \frac{(n-2) \Gamma(n-1) (1-\rho^2)^{\frac{n-1}{2}} (1-r^2)^{\frac{n-4}{2}}}{\sqrt{2\pi} \Gamma(n-\frac{1}{2}) (1-\rho r)^{n-\frac{3}{2}}} {}_2F_1\left(\frac{1}{2}, \frac{1}{2}; \frac{2n-1}{2}; \frac{\rho r+1}{2}\right)$$

其中 Γ 是伽玛函数, ${}_2F_1(a, b; c; z)$ 是高斯超几何函数。

注意到 $E(r) = \rho - \frac{\rho(1-\rho^2)}{2(n-1)} + \dots$, 因此 r 是 ρ 的一个有偏估计。一种获得无偏估计的方法是解 ρ 的方程 $r = E(r) = \rho - \frac{\rho(1-\rho^2)}{2(n-1)}$ 。然而, 解 $\check{\rho} = r \left[1 + \frac{1-r^2}{2(n-1)} \right]$ 是次优的。一种无偏估计, 可以从 n 较大情况下的最小方差和有偏序列 $\frac{1}{n-1}$, 通过最大化 $\log f(r)$, 也就是 $\hat{\rho} = r \left[1 - \frac{1-r^2}{2(n-1)} \right]$ 获得。

特殊情况下, 当 $\rho = 0$ 时, 分布可以被写成

$$f(r) = \frac{(1-r^2)^{\frac{n-4}{2}}}{B\left(\frac{1}{2}, \frac{n-2}{2}\right)}$$

其中 B 是贝塔函数。

费舍尔变换

实际应用中, 与 ρ 相关的置信区间和假设检验通常是通过费舍尔变换获得

$$F(r) = \frac{1}{2} \ln \frac{1+r}{1-r} = \operatorname{arctanh}(r).$$

如果 $F(r)$ 是 r 的费舍尔变换, n 是样本容量, 那么 $F(r)$ 近似服从正态分布

$$\text{mean} = F(\rho) = \operatorname{arctanh}(\rho) \quad \text{and standard error} \quad \text{SE} = \frac{1}{\sqrt{n-3}}.$$

也就是标准分是

$$z = \frac{x - \text{mean}}{\text{SE}} = [F(r) - F(\rho_0)]\sqrt{n-3}$$

对 $\rho = \rho_0$ 进行零假设, 可以设想样本数据对是独立同分布并且服从双变量正态分布。因此P值估计可以从正态分布概率表中获得。比如, 如果观测数据 $z = 2.2$, 并且要用双边p值对 $\rho = 0$ 进行零假设检验, p值是 $2 \cdot \Phi(-2.2) = 0.028$, 其中 Φ 是正态分布的累积分布函数。

置信区间

为了获得 ρ 的信赖区间, 首先, 我们应该计算 $F(\rho)$ 的信赖区间:

$$100(1 - \alpha)\%CI : \operatorname{arctanh}(\rho) \in [\operatorname{arctanh}(r) \pm z_{\alpha/2}SE]$$

通过可逆Fisher变换可以获得相关尺度上的区间。

$$100(1 - \alpha)\%CI : \rho \in [\tanh(\operatorname{arctanh}(r) - z_{\alpha/2}SE), \tanh(\operatorname{arctanh}(r) + z_{\alpha/2}SE)]$$

举例来说，假设我们观测到 $r = 0.3$ ，样本容量 $n=50$ ，并且我们期望获得 ρ 的 95% 的信赖区间。变换后的值是 $\operatorname{artanh}(r) = 0.30952$ ，所以在变换尺度上的信赖区间是 $0.30952 \pm 1.96/\sqrt{47}$ ，或者 $(0.023624, 0.595415)$ 。变换回相关尺度上是 $(0.024, 0.534)$ 。

皮尔逊相关系数和最小方差回归分析

样本相关系数的平方，亦称作 coefficient of determination, 利用简单线性回归估计由 X 引起的 Y 的变化。一开始, Y_i 围绕它们平均值上的变化可以分解成

$$\sum_i (Y_i - \bar{Y})^2 = \sum_i (Y_i - \hat{Y}_i)^2 + \sum_i (\hat{Y}_i - \bar{Y})^2,$$

其中 \hat{Y}_i 是作回归分析时的适应值。整理后得

$$1 = \frac{\sum_i (Y_i - \hat{Y}_i)^2}{\sum_i (Y_i - \bar{Y})^2} + \frac{\sum_i (\hat{Y}_i - \bar{Y})^2}{\sum_i (Y_i - \bar{Y})^2}.$$

两个被加数是由 X (右边) 引起的 Y 的变化和不是由 X (左边) 引起的变化。

接下来, 我们利用最小方差回归模型, 使 \hat{Y}_i 和 $Y_i - \hat{Y}_i$ 的样本协方差为 0。于是, 观测数据和适应值的样本相关系数可以被写成

$$\begin{aligned} r(Y, \hat{Y}) &= \frac{\sum_i (Y_i - \bar{Y})(\hat{Y}_i - \bar{Y})}{\sqrt{\sum_i (Y_i - \bar{Y})^2 \cdot \sum_i (\hat{Y}_i - \bar{Y})^2}} \\ &= \frac{\sum_i (Y_i - \hat{Y}_i + \hat{Y}_i - \bar{Y})(\hat{Y}_i - \bar{Y})}{\sqrt{\sum_i (Y_i - \bar{Y})^2 \cdot \sum_i (\hat{Y}_i - \bar{Y})^2}} \\ &= \frac{\sum_i [(Y_i - \hat{Y}_i)(\hat{Y}_i - \bar{Y}) + (\hat{Y}_i - \bar{Y})^2]}{\sqrt{\sum_i (Y_i - \bar{Y})^2 \cdot \sum_i (\hat{Y}_i - \bar{Y})^2}} \\ &= \frac{\sum_i (\hat{Y}_i - \bar{Y})^2}{\sqrt{\sum_i (Y_i - \bar{Y})^2 \cdot \sum_i (\hat{Y}_i - \bar{Y})^2}} \\ &= \sqrt{\frac{\sum_i (\hat{Y}_i - \bar{Y})^2}{\sum_i (Y_i - \bar{Y})^2}}. \end{aligned}$$

于是

$$r(Y, \hat{Y})^2 = \frac{\sum_i (\hat{Y}_i - \bar{Y})^2}{\sum_i (Y_i - \bar{Y})^2}$$

是由 X 的线性方程引起的 Y 的平均变化。

数据分布的敏感度

存在性

总体皮尔逊相关系数被定义成矩, 因此任意的双变量概率分布是非零的, 也就是说 总体协方差 和 边缘 总体方差 是由定义的。一些概率分布, 诸如 柯西分布 有未定义的方差, 因此 X or Y 如果服从这种分布, ρ 便是未定义的。在实际应用中, 如果有数据被怀疑服从重尾分布, 这个条件就需要引起重视。然而, 相关系数的存在性通常并不需要太介意; 例如, 如果分布是有界的, ρ 便总是有意义的。

大样本的特性

在双变量 正态分布 的案例中, 只要边缘均值和方差是已知的, 总体相关系数描述的是便是联合分布。在其他的双变量分布中, 这个结论并不正确。总之, 不论两个随机变量的联合分布是不是正态的, 相关系数在研究的它们之间的线性依赖性都是有帮助的。^[2] 样本相关系数是对两个正态分布变量总体相关系数的最大似然估计 并且是 渐进 无偏的 和 有效的, 这也就是说如果数据是正态的并且样本容量是中等的或大量的, 就不可能构造出一个比样本相关系数更准确的估计。对于非正态的数据, 样本相关系数大致上是无偏的, 但有可能是无效的。只要样本均值、方差和协方差是一致的 (当大数定理可以应用的情况下), 样本相关系数是总体相关系数的一致估计。

稳健性

与其他常用的统计指标相似的, 样本指标 r 不是 稳健的^[15]。因此如果由 异常值, 这个指标是有误导性的。^{[16][17]} 特别的, PMCC 既不是稳健分布的, 也不是异常值稳健的^[15] (see Robust statistics#Definition)。对 X 和 Y 的散点图的观察可以很明显的揭示出缺乏稳健性的情况, 在这种情况下, 采用的联合的方法是比较明智的。注意到, 虽然大多数稳健的估计器从某种程度上说都是有统计依赖的, 它们总的来说, 在总体相关系数的尺度上都是可辨的。

基于皮尔逊相关系数的统计推断对数据分布式敏感的。如果数据大致是正态分布的, 可以使用精确检验和基于Fisher变换的渐进检验, 但是它们可能由误导性。在一些情况下, 自助采样 可以用来构造置信区间。同时, 重复抽样 可以应用在假设检验中。这些非参数化的方法在某些情况下, 如双变量正态分布不能保证时, 可能得出更有意义的结论。然而, 这些方法的标准形式依赖于数据的 可交换性。这也就意味着被分析的数据时没有顺序的和组别的。因为这有可能会影响估计相关系数的特性。

分层分析是一种容许缺少双变量正态性的方法, 或者说是用来隔离相互关联因素的关联结果。如果 W 代表聚类成员或者其它需要被控制的因素, 我们可以分离基于 W 的数据, 然后我们可以再每个层里计算相关系数。当我们控制变量 W , 我们便能在层的等级上估计与所有相关系数相关的各自的相关系数。^[18]

计算加权相关系数

假设我们要计算关联性的观测数据有着不同的重要程度, 表示成权值向量 w 。利用权值向量 w (总长度 n) 计算向量 x 和 y 的相关系数,^[19]

- 加权均值:

$$\mathbf{m}(x; w) = \frac{\sum_i w_i x_i}{\sum_i w_i}.$$

- 加权协方差

$$\text{cov}(x, y; w) = \frac{\sum_i w_i (x_i - \mathbf{m}(x; w))(y_i - \mathbf{m}(y; w))}{\sum_i w_i}.$$

- 加权相关系数

$$\text{corr}(x, y; w) = \frac{\text{cov}(x, y; w)}{\sqrt{\text{cov}(x, x; w) \text{cov}(y, y; w)}}.$$

去除相关性

我们总是可以通过一定的线性变换去除随机变量之间的相关性,即便变量间的关系是非线性的。Cox & Hinkley^[20]给出了在总体相关系数中的表达形式。

与此相应的,样本相关系数也存在这样的结论,使得样本相关系数变为0。假设长度为 n 的随机变量被随机采样 m 次。令 X 是一个矩阵,其中 $X_{i,j}$ 是第 i 次采样的第 j 个变量。令 $Z_{m,m}$ 是一个所有元素都为1的 $m * m$ 的方阵。那么 D 是变换后的数据,使得随机变量的均值为0,并且 T 是变换后的数据,使得所有的变量均值为0和与除自身外的其他变量的相关系数为0 - T 的矩作为身份矩阵。为了得到单位方差,还需要除以标准差。虽然变换后的数据有可能不是独立的,但他们一定是不相关的。

$$D = X - \frac{1}{m} Z_{m,m} X$$

$$T = D(D^T D)^{-\frac{1}{2}}$$

其中,指数-1/2表示矩阵置换后的 矩阵方根。 T 的协方差被当做身份矩阵。如果新的样本数据 x 是 n 个元素的向量,那么相同的变换可以应用到 x 中以获得变换向量 d 和 t :

$$d = x - \frac{1}{m} Z_{1,m} X$$

$$t = d(D^T D)^{-\frac{1}{2}}$$

这个去相关性的方法被应用到多变量的主成分分析中。

反射相关性

反射相关系数是皮尔逊相关系数的变体,数据并不是以他们的均值为中心。总体反射相关系数是

$$\text{Corr}_r(X, Y) = \frac{E[XY]}{\sqrt{EX^2 \cdot EY^2}}.$$

反射相关系数是对称的,但在如下的变换中并不是不变的

$$\text{Corr}_r(X, Y) = \text{Corr}_r(Y, X) = \text{Corr}_r(X, bY) \neq \text{Corr}_r(X, a + bY), \quad a \neq 0, b > 0.$$

样本反射相关系数是

$$rr_{xy} = \frac{\sum x_i y_i}{\sqrt{(\sum x_i^2)(\sum y_i^2)}}.$$

样本加权相关系数是

$$rr_{xy,w} = \frac{\sum w_i x_i y_i}{\sqrt{(\sum w_i x_i^2)(\sum w_i y_i^2)}}.$$

比例关系

规模的相关性是一个变种的皮尔森相关数据的范围限制故意以受控的方式揭示时间序列之间的快速成分的相关性。比例相关的定义是在短数据段的平均相关性。对于给定规模S，令K为可以适应信号的总长度的段数：

$$\mathbf{K} = \mathbf{Sound} \left(\frac{T}{s} \right)$$

比例相关的整个信号的 r_s 的计算公式为

$$\vec{r_s} = \frac{1}{K} \sum_{k=1}^K r_k$$

r_s 为k的部分皮尔森相关系数。通过对参数s的选择，减少值的范围和较长的时间尺度上的相关性被过滤掉，只有在很短的时间尺度上的相关性被发现。因此，慢分量的贡献被删除，快分量被保留。

强噪声条件下

强噪声条件下，提取相关系数两个随机变量之间的是平凡的，特别是在典型相关分析报告在退化的相关值的情况下，由于存在大量噪声。一种概括的方法在其他地方给出。

维基相关条目

- 相关
- 史匹曼等级相关系数
- Disattenuation
- Maximal information coefficient
- 图模式
- 马尔可夫链
- 马尔可夫逻辑网络

外部链接

- 相关系数：使用皮尔森检定与斯皮尔曼等级检定的研究问题 (<http://mail.cmu.edu.tw/~tcli/%B2%C4%A4Q%A4%BB%B3%B9.pdf>) (中国医药大学，生物统计课程)
- 相关系数：资料型态与适用统计方法 (http://www2.cmu.edu.tw/~biostat/online/teaching_corner_038-2.pdf) (中国医药大学，生物统计课程)

注脚

1. "The human disease network", Albert Barabasi et al., Plos.org
2. J. L. Rodgers and W. A. Nicewander. Thirteen ways to look at the correlation coefficient (<http://www.jstor.org/stable/2685263>). The American Statistician, 42(1):59–66, February 1988.
3. Stigler, Stephen M. Francis Galton's Account of the Invention of Correlation. Statistical Science. 1989, 4 (2): 73–79. JSTOR 2245329. doi:10.1214/ss/1177012580.
4. A. Buda and A.Jarynowski (2010) *Life-time of correlations and its applications vol.1*, Wydawnictwo Niezalezne: 5–21, December 2010, ISBN 978–83–915272–9–0
5. Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (2nd ed.)
6. Fulekar (Ed.), M.H. (2009) *Bioinformatics: Applications in Life and Environmental Sciences*, Springer (pp. 110) ISBN 1402088795

7. N.A Rahman, A Course in Theoretical Statistics; Charles Griffin and Company, 1968
8. Kendall, M.G., Stuart, A. (1973) *The Advanced Theory of Statistics, Volume 2: Inference and Relationship*, Griffin. ISBN 0852642156 (Section 31.19)
9. Fisher, R.A. Frequency distribution of the values of the correlation coefficient in samples from an indefinitely large population. *Biometrika*. 1915, **10** (4): 507–521. doi:10.1093/biomet/10.4.507.
10. Fisher, R.A. On the probable error of a coefficient of correlation deduced from a small sample (PDF). *Metron*. 1921, **1** (4): 3–32 [2009–03–25].
11. Gayen, A.K. The frequency distribution of the product moment correlation coefficient in random samples of any size draw from non-normal universes. *Biometrika*. 1951, **38**: 219–247. doi:10.1093/biomet/38.1–2.219.
12. Soper, H.E., Young, A.W., Cave, B.M., Lee, A., Pearson, K. (1917). "On the distribution of the correlation coefficient in small samples. Appendix II to the papers of "Student" and R. A. Fisher. A co-operative study", *Biometrika*, **11**, 328–413. doi:10.1093/biomet/11.4.328
13. Kenney, J. F. and Keeping, E. S., *Mathematics of Statistics*, Pt. 2, 2nd ed. Princeton, NJ: Van Nostrand, 1951.
14. Correlation Coefficient – Bivariate Normal Distribution (<http://mathworld.wolfram.com/CorrelationCoefficientBivariateNormalDistribution.html>)
15. Wilcox, Rand R. Introduction to robust estimation and hypothesis testing. Academic Press. 2005.
16. Devlin, Susan J; Gnanadesikan, R; Kettenring J.R. Robust Estimation and Outlier Detection with Correlation Coefficients. *Biometrika*. 1975, **62** (3): 531–545. JSTOR 2335508. doi:10.1093/biomet/62.3.531.
17. Huber, Peter. J. Robust Statistics. Wiley. 2004.
18. Katz., Mitchell H. (2006) *Multivariable Analysis – A Practical Guide for Clinicians*. 2nd Edition. Cambridge University Press. ISBN 9780521549851. ISBN 052154985X doi:10.2277/052154985X (<http://dx.doi.org/10.2277%2F052154985X>)
19. <http://sci.tech-archive.net/Archive/sci.stat.math/2006-02/msg00171.html></ref><ref>A MATLAB Toolbox for computing Weighted Correlation Coefficients (<http://www.mathworks.com/matlabcentral/fileexchange/20846>)
20. Cox, D.R., Hinkley, D.V. (1974) *Theoretical Statistics*, Chapman & Hall (Appendix 3) ISBN 0412124203

取自“<https://zh.wikipedia.org/w/index.php?title=皮尔逊积矩相关系数&oldid=56568178>”

本页面最后修订于2019年10月22日 (星期二) 02:15。

本站的全部文字在知识共享 署名–相同方式共享 3.0协议之条款下提供，附加条款亦可能应用。（请参阅使用条款）
Wikipedia®和维基百科标志是维基媒体基金会的注册商标；维基™是维基媒体基金会的商标。
维基媒体基金会是按美国国内税收法501(c)(3)登记的非营利慈善机构。