全国中文核心期刊（2004版）

2017 3

大学数学

College Mathematics

Daxue Shuxue

第三十三卷 VOL. 33

ISSN 1672-1454

大 学 数 学

2017 年 6 月　　　　第 33 卷　第 3 期(总第 191 期)

目　录

[期刊基本参数]　CN 34－1221/O1＊1984＊b＊A4＊128＊zh＋en＊P＊￥15.00＊1200＊24＊2017－06

本期责任编辑：孙　琳，周　玲，涂振坤

第 33 卷第 3 期　　大 学 数 学　　Vol. 33, №. 3
2017 年 6 月　　COLLEGE MATHEMATICS　　Jun. 2017

相关系数概念剖析

崔艳丽

(烟台大学 数学与信息科学学院，山东 烟台 264005)

［摘　要］从两个角度——内积空间以及线性回归角度深入剖析了相关系数这一重要概念，将其与 $\mathbb{R}^2$ 空间中向量之间的夹角联系起来，并且给出了一种迅速判断随机变量之间相关性强弱的方法，并通过随机模拟进行了直观展示.

［关键词］同构；夹角；随机模拟

［中图分类号］O211.1　　［文献标识码］C　　［文章编号］1672-1454(2017)03-0114-04

1　引　　言

相关系数是描述两个随机变量之间线性相关关系强弱的重要指标，但是在一般教科书中并未对此概念进行较为深入的探讨，不利于教学过程中学生的理解. 本文尝试从两个角度较为透彻地剖析了这个基本概念，并且给出了一种直观判断随机变量之间线性相关程度强弱的方法，然后通过随机模拟进行了展示.

2　从内积角度理解相关系数

定义 1[1]　设 (Ω,F,P) 是概率空间，C 是定义在 Ω 上的二阶矩有穷的随机变量的全体，即

$$C=\left\{X\left|\mathrm{E}X^2=\int_\Omega X^2(\omega)P(\mathrm{d}\omega)<\infty\right.\right\},$$

容易验证 C 是线性空间. 对任意的 $X,Y\in C$，定义 $(X,Y)=E(XY)$，容易验证其为内积，这个空间称为 $L^2(\Omega,F,P)$，简称 L^2 空间.

定义 2[1]　实内积空间中任意两个非零元素 x,y 之间的夹角定义为

$$\theta=\arccos\frac{(x,y)}{\|x\|\ \|y\|}.$$

特别地，L^2 空间中任意两个非零元素 X,Y 的夹角

$$\theta=\arccos\frac{\mathrm{E}(XY)}{\sqrt{\mathrm{E}X^2}\ \sqrt{\mathrm{E}Y^2}}.$$

在以上知识的基础上，给出相关系数的定义.

定义 3　设 $X,Y\in L^2$，且 $\mathrm{E}X=\mu_1,\mathrm{E}Y=\mu_2,DX=\sigma_1^2\neq0,DY=\sigma_2^2\neq0$，将 X,Y 标准化，不妨令

$$X^*=\frac{X-\mu_1}{\sigma_1},\quad Y^*=\frac{X-\mu_2}{\sigma_2},$$

X^* 与 Y^* 的夹角记为 $\theta(0\leqslant\theta\leqslant\pi)$，称 $\cos\theta$ 为 X,Y 的相关系数，记为 ρ. 补充定义常数与任何随机变

［收稿日期］2016-09-09；［修改日期］2017-03-22

［基金项目］博士启动基金(SX09B23)，烟台大学数学与信息科学学院教改项目

［作者简介］崔艳丽(1981－)，女，博士，讲师，从事数理统计研究. Email:cuiyanli@amss.ac.cn

量的相关系数为 0.

定理 1　设 M 为 X^* 与 Y^* 张成的闭线性子空间,则 M 是可分的 Hilbert 空间.

证　$$\begin{aligned} M &= \overline{sp}\{X^*, Y^*\} = \{Z : Z = aX^* + bY^*, a, b \in \mathbb{R}\} \\ &= \{Z : Z = aX^* + b(Y^* - \rho X^*), a, b \in \mathbb{R}\} \\ &= \overline{sp}\{X^*, Y^* - \rho X^*\}, \end{aligned}$$

易知 X^* 与 $Y^* - \rho X^*$ 正交,因此 M 是可分的 Hilbert 空间.

定理 2　M 与 $\mathbb{R}^2$ 同构.

证　考虑映射:$Th \to ((h, X^*), (h, Y^* - \rho X^*))$,下面证明 T 是 $1-1$ 映射.

先证 T 是单射. $\forall h_1, h_2 \in M$,且 $h_1 \neq h_2$,由于 M 可分,X^* 与 $\dfrac{Y^* - \rho X^*}{\sqrt{1-\rho^2}}$ 为其一组标准正交基,不妨设

$$h_1 = a_1 X^* + b_1 \frac{(Y^* - \rho X^*)}{\sqrt{1-\rho^2}},$$

$$h_2 = a_2 X^* + b_2 \frac{(Y^* - \rho X^*)}{\sqrt{1-\rho^2}}, \quad (a_1, b_1) \neq (a_2, b_2).$$

由定义知

$$Th_1 = ((h_1, X^*), (h_1, Y^* - \rho X^*)) = (a_1, b_1),$$

$$Th_2 = ((h_2, X^*), (h_2, Y^* - \rho X^*)) = (a_2, b_2),$$

故 $Th_1 \neq Th_2$.

下证 T 是满射. $\forall (a, b) \in \mathbb{R}^2$,令

$$h = aX^* + b\frac{(Y^* - \rho X^*)}{\sqrt{1-\rho^2}},$$

则 $Th = (a, b)$.

综上 T 是 $1-1$ 映射,M 与 $\mathbb{R}^2$ 同构.

定理 3　设 X^* 与 Y^* 在 L^2 中的夹角为 θ_1,$(1,0)$ 与 $(\rho, \sqrt{1-\rho^2})$ 在 $\mathbb{R}^2$ 中的夹角为 θ_2,则 $\theta_1 = \theta_2$.

证　根据 $\mathbb{R}^2$ 中内积的定义

$$\cos\theta_2 = \frac{1 \times \rho + 0 \times \sqrt{1-\rho^2}}{\sqrt{1} \times \sqrt{1}} = \rho = \cos\theta_1,$$

因此 $\theta_1 = \theta_2$.

上面的定理将随机变量间的相关系数 ρ 与我们熟悉的 $\mathbb{R}^2$ 空间中向量的夹角联系起来,从而将抽象的概念直观化,具体参见图 1.

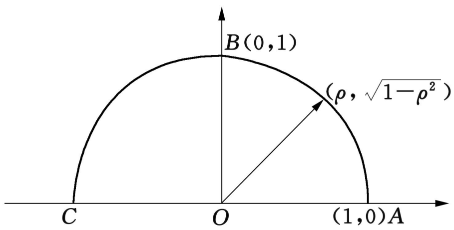

图 1　相关系数与夹角同步变化示意图

可见,当 ρ 从 0 到 1 时,$(\rho, \sqrt{1-\rho^2})$ 从 B 点沿圆周滑到 A 点,对应向量之间的夹角由 $\dfrac{\pi}{2}$ 缩小到 0,向量 $(\rho, \sqrt{1-\rho^2})$ 在 $(1,0)$ 方向上的投影逐渐变长且与之同向,显示越来越强的正相关性;当 ρ 从 0 到 -1 时,$(\rho, \sqrt{1-\rho^2})$ 从 B 点沿圆周滑到 C 点,对应向量之间的夹角由 $\dfrac{\pi}{2}$ 增大到 π,向量 $(\rho, \sqrt{1-\rho^2})$ 在 $(1,0)$ 方向上的投影逐渐变长但与之反向,显示越来越强的负相关性.

3　从回归角度理解相关系数及相关性强弱的几何直观化

3.1　从回归角度理解相关系数

既然我们关心两随机变量 Y, X 的线性关系强弱,不妨将 Y 关于 X 作线性回归,同时为了消除量纲

的影响,将二者先进行标准化处理. 结果如下:

定理4 设 $X,Y\in L^2$,且 $EX=\mu_1,EY=\mu_2,DX=\sigma_1^2\neq 0,DY=\sigma_2^2\neq 0$,将 X,Y 标准化,不妨令 $X^*=\dfrac{X-\mu_1}{\sigma_1}$,$Y^*=\dfrac{X-\mu_2}{\sigma_2}$,则 Y^* 关于 X^* 的最佳线性预测为 $h(X^*)=\rho X^*$,即

$$\| Y^*-h(X^*)\|^2=\min_{a,b\in\mathbb{R}}\| Y^*-aX^*-b\|^2$$

证
$$\begin{aligned}\| Y^*-aX^*-b\|^2&=E(Y^*-aX^*-b)^2=E[(Y^*-aX^*)^2-2b(Y^*-aX^*)+b^2]\\&=EY^{*2}+a^2EX^{*2}-2aEX^*Y^*+b^2=1+a^2-2a\rho+b^2\\&=(a-\rho)^2+b^2+1-\rho^2.\end{aligned}$$

显然当 $a=\rho,b=0$ 时,$\| Y^*-aX^*-b\|^2$ 达到最小,且最小值为 $1-\rho^2$.

由上述定理知道 Y^* 关于 X^* 的最佳线性预测仅依赖于 ρ,ρ 为正说明二者具有正向的相关关系,ρ 为负说明二者具有负向的相关关系,并且这种关系随 $|\rho|$ 增大而增强,越接近1,预测效果越好,当 $|\rho|=1$ 时,可以完全准确地进行线性预测. 因此 ρ 可以作为衡量两个变量线性相关程度的指标,故称线性相关系数或相关系数.

3.2 相关性强弱的几何直观化

受上述定理启发,我们可以先对标准化后的数据进行线性拟合,然后通过观察所得拟合直线的斜率,对随机变量的相关性强弱作出判断. 直线通过一、三象限且越接近 x 轴,即斜率越接近0,表示相关性越小;越接近 $y=x$,即斜率越接近1,表示正相关性越强;直线通过二、四象限且越接近 $y=-x$,即斜率越接近-1,表示负相关性越强. 下面我们产生一组来自正态分布的随机数,在不同的 ρ 值下进行直观展示.

首先产生来自二元正态分布 $N(\mathbf{0},\boldsymbol{I}_{2\times 2})$ 的一组随机数 $\begin{pmatrix}D_1\\D_2\end{pmatrix}_{2\times 30}$,则由多元正态分布性质易知,

$$\begin{pmatrix}D_1^*\\D_2^*\end{pmatrix}=\begin{pmatrix}\sigma_1^2&\rho\sigma_1\sigma_2\\\rho\sigma_1\sigma_2&\sigma_2^2\end{pmatrix}^{\frac{1}{2}}\begin{pmatrix}D_1\\D_2\end{pmatrix}+\begin{pmatrix}\mu_1 1_{1\times 30}\\\mu_2 1_{1\times 30}\end{pmatrix}$$

是来自正态分布 $N\left(\begin{pmatrix}\mu_1\\\mu_2\end{pmatrix},\begin{pmatrix}\sigma_1^2&\rho\sigma_1\sigma_2\\\rho\sigma_1\sigma_2&\sigma_2^2\end{pmatrix}\right)$ 的一组随机数,并且 D_1^* 与 D_2^* 是分别来自一元正态分布 $N(\mu_1,\sigma_1^2)$ 和 $N(\mu_2,\sigma_2^2)$ 的随机数,相关系数为 ρ. 以 D_{1i}^* 为横坐标,D_{2i}^* 为纵坐标,得到坐标点 (D_{1i}^*,D_{2i}^*),$i=1,\cdots,30$. 不妨取 $\sigma_1^2=1,\sigma_2^2=4,\mu_1=1,\mu_2=2$,根据 ρ 值的不同,分以下情况讨论:

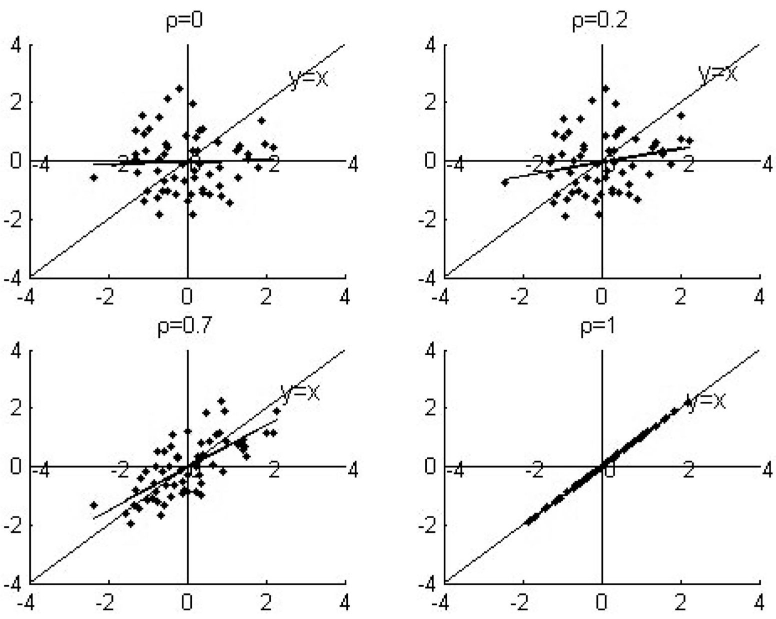

图2 相关性强弱直观判断示意图

(a) $\rho = 0$

从图中可以看出,经线性拟合后所得直线几乎与 x 轴重合,说明随机变量间不存在相关关系.

(b) $0 < \rho < 0.5$,不妨取 $\rho = 0.2$

经线性拟合后所得直线的斜率为正,但与 $y = x$ 的距离较远,说明随机变量间存在正相关关系,但相关程度较弱.

(c) $0.5 < \rho < 1$,不妨取 $\rho = 0.7$

经线性拟合后所得直线的斜率为正,并且与 $y = x$ 的距离较近,说明随机变量间存在正相关关系,并且相关程度较强.

(d) $\rho = 1$

经线性拟合后所得直线几乎与 $y = x$ 重合,说明随机变量间完全正相关.

注　当 $-1 < \rho < 0$ 时,通过上述类似的讨论可以得到相似的结论.

[参　考　文　献]

[1] 田铮译. 时间序列的理论与方法[M]. 北京:高等教育出版社,2011.

[2] 李贤平. 概率论基础[M]. 北京:高等教育出版社,2010.

[3] 王学民. 应用多元分析[M]. 上海:上海财经大学出版社,2010.

Two Ways of Understanding the Coefficient of Correlation

CUI Yan-li

(School of Mathematics and Information Science, Yantai University, Yantai Shandong 264005, China)

Abstract: We discussed two ways of understanding the most important concept in probability——the coefficient of correlation. We found it was related to the angle between two vectors in $\mathbb{R}^2$ and also proposed a approach through which we could judge the correlation strength immediately and visually.

Key words: isomorphism; angle; random simulation